

Improved approaches to ligand growing through fragment docking and fragment-based library design

Dissertation

zur Erlangung des Doktorgrades
der Naturwissenschaften
(Dr. rer. nat.)

dem
Fachbereich Pharmazie der
Philipps-Universität Marburg
vorgelegt von

Florent Chevillard

aus
Toulouse

Marburg/Lahn 2016

Erstgutachter Dr. Peter Kolb
 Institut für Pharmazeutische Chemie
 Philipps-Universität Marburg

Zweitgutachter Prof. Dr. Gerhard Klebe
 Institut für Pharmazeutische Chemie
 Philipps-Universität Marburg

Eingereicht am 28.7.2016

Tag der mündlichen Prüfung am 23.9.2016

Hochschulkennziffer: 1180

"And why do we fall, Bruce ? So we can learn to pick ourselves up."

Thomas Wayne, *Batman begins* (2005)

Abstract

In the past two decades, fragment-based drug discovery (FBDD) has continuously gained popularity in drug discovery efforts and has become a dominant tool in order to explore novel chemical entities that might act as bioactive modulators. FBDD is intimately connected to fragment extension approaches, such as growing, merging or linking. These approaches can be accelerated using computational programs or semi-automated workflows for *de novo* design. Although computers allow for the facile generation of millions of suggestions, this often comes at a price: uncertain synthetic feasibility of the generated compounds, potentially leading to a dead end in an optimization process.

In this manuscript we developed two computational tools which could support the FBDD elaboration cycle: PINGUI and SCUBIDOO. PINGUI is a semi-automated workflow for fragments growing guided by both the protein structure and synthetic feasibility. SCUBIDOO is a freely accessible database which currently holds 21 M virtual products. This database was created by combining commercially available building blocks with robust organic reactions. Thus, every virtual product comes with synthetic instructions. Most of the crucial functions of PINGUI (creation of derived libraries or applying organic reaction) were then implemented in the SCUBIDOO website.

PINGUI and SCUBIDOO were then applied to fragment-based ligand discovery efforts targeting the β_2 -adrenergic receptor (β_2 AR) and the PIM1 kinase. In a first study focusing on the β_2 AR, we suggested a total of eight diverse extensions for different fragment hits using PINGUI. The eight compounds were successfully synthesized and further assays showed that four products had an improved affinity compared to their respective initial fragment. In a second study, SCUBIDOO was applied in order to quickly identify fragments and suggest extensions that could bind to PIM1. This study yielded a fragment hit and its associated crystal structure. Synthesis of derived products is in progress. Lastly, SCUBIDOO was coupled with automated robotic synthesis in order to synthesize hundreds of compounds in parallel. 127 products among the 240 suggested were synthesized (53%). Those compounds were designed so they are likely to bind to the β_2 AR and will be tested in the near future.

The aforementioned computational tools could improve early fragment-based drug discovery projects, especially in the realm of fragment growing strategies. For instance, PINGUI suggests extensions that are very likely to be attachable, making it a useful creative tool for medicinal chemists during structure-activity relationship (SAR) studies. With so far 53% success synthesis rate, SCUBIDOO has shown that it is amenable to be integrated to automated robotic synthesis. Every synthesis attempt is prone to

improve the knowledge contained within the database and thus increase the synthesis success rate over time. Furthermore, all synthesized product were novel compounds, thus demonstrating how SCUBIDOO could explore new quadrants of the chemical space.

Zusammenfassung

Die Fragment-basierte Wirkstoffforschung ("fragment-based drug discovery" – FBDD) hat in den vergangenen zwei Jahrzehnten kontinuierlich an Beliebtheit gewonnen und sich zu einem dominanten Instrument der Erforschung neuer chemischer Moleküle als potentielle bioaktive Modulatoren entwickelt. FBDD ist eng mit Ansätzen zur Fragment-Erweiterung, wie etwa dem Fragment-"growing", "merging" oder dem "linking", verknüpft. Diese Entwicklungsansätze können mit Hilfe von Computerprogrammen oder teilautomatischen Prozessen der "de novo" Wirkstoffentwicklung beschleunigt werden. Obwohl Computer mühelos Millionen von Vorschlägen generieren können, geschieht dies allerdings oft auf Kosten unsicherer synthetischer Realisierbarkeit der Verbindungen mit einer potentiellen Sackgasse im Optimierungsprozess.

Dieses Manuskript beschreibt die Entwicklung zweier computerbasierter Instrumente, PINGUI und SCUBIDOO, mit dem Ziel den FBDD Ausarbeitungs-Zyklus zu fördern. PINGUI ist ein halbautomatischer Arbeitsablauf zur Fragment-Erweiterung basierend auf der Proteinstruktur unter Berücksichtigung der synthetischen Umsetzbarkeit. SCUBIDOO ist eine freizugängliche Datenbank mit aktuell 21 Millionen virtuellen Produkten, entwickelt durch die Kombination kommerziell verfügbarer Bausteine ("building blocks") mit bewährten organischen Reaktionen. Zu jedem erzeugten virtuellen Produkt wird somit eine Synthesevorschrift geliefert. Die entscheidenden Funktionen von PINGUI, wie die Erzeugung abgeleiteter Bibliotheken oder das Anwenden organischer Reaktionen, wurden daraufhin in die SCUBIDOO Webseite integriert.

PINGUI als auch SCUBIDOO wurden des Weiteren zur Erforschung Fragment-basierter Liganden ("fragment-based ligand discovery") mit dem β_2 -adrenergen Rezeptor (β_2 AR) und der PIM1 Kinase als Zielproteine eingesetzt. Im Rahmen einer ersten Studie zum β_2 AR wurden mit PINGUI acht unterschiedliche Erweiterungen für verschiedene Fragment-Treffer ("hits") vorhergesagt. Alle acht Verbindungen konnten dabei erfolgreich synthetisiert werden und vier der acht Produkte zeigten im Vergleich zu den Ausgangsfragmenten eine erhöhte Affinität zum target. Eine zweite Studie umfasste die Anwendung von SCUBIDOO zur schnellen Identifikation von Fragmenten und deren möglichen Erweiterungen mit potentieller Bindungsaktivität zur PIM1 Kinase. Als Ergebnis ergab sich ein Fragment-Treffer mit der dazugehörigen Kristallstruktur. Weitere Folgeprodukte befinden sich derzeit in Synthese. Abschließend wurde SCUBIDOO an eine automatische Roboter- Synthese gekoppelt, wodurch hunderte von Verbindungen effizient parallel synthetisiert werden können. 127 der 240 vorhergesagten Produkte (53%) wurden mit dem Ziel an den β_2 AR zu binden bereits synthetisiert und werden in Kürze weitergehend getestet.

Die beiden vorgestellten Computer-Tools könnten zur Verbesserung im Anfangsstadium befindlicher Projekte zur Fragment-basierten Wirkstoffentwicklung, vor allem hinsichtlich der Strategien im Bereich der Fragment Erweiterung, eingesetzt werden. PIN-GUI zum Beispiel generiert Vorschläge zur Fragment- Erweiterung, die sich mit hoher Wahrscheinlichkeit an die Zielstruktur anlagern, und stellt somit ein nützliches und kreatives Werkzeug zur Untersuchung von Struktur-Wirkungsbeziehungen ("structure-activity relationship" – SAR) dar. SCUBIDOO zeigte sich mit einem bisherigen 53-prozentigen Synthese-Erfolg als zugänglich für die Integration an die effiziente automatisierte Roboter-Synthese. Jede zukünftige Synthese liefert neue Kenntnisse innerhalb der Datenbank und wird somit nach und nach den Synthese-Erfolg erhöhen. Des Weiteren stellen alle synthetisierten Produkte neuartige Verbindungen dar, was umso mehr den möglichen Einfluss SCUBIDOOs bei der Entdeckung neuer chemischer Strukturen hervorhebt.

Contents

Abstract	iii
Contents	vii
List of Figures	xii
List of Tables	xiv
Abbreviations	xv
1 Introduction	1
2 Fragment Based Drug Discovery	7
2.1 Introduction	7
2.2 Fragment libraries	11
2.2.1 Important properties for experimental screening	11
2.2.1.1 Library size	11
2.2.1.2 Chemical diversity	11
2.2.1.3 Solubility	12
2.2.1.4 Reactive functional groups	12
2.2.1.5 Rule of three compatibility	12
2.2.1.6 Specialized application	13
2.2.2 Important properties for virtual screening	13
2.3 In Silico Approaches	14
2.3.1 Structure-based methods	15
2.3.1.1 Defining the binding site	15
2.3.1.2 Exploring the conformational space of the ligand	15
Exhaustive search.	15
Stochastic search.	16
2.3.1.3 Orientational sampling	16
Systematic and random searches.	16
Shape complementarity methods.	17
Point complementarity methods.	17
Incremental methods.	18
2.3.1.4 Scoring	18
Force-field-based.	18
Empirical.	20

	Knowledge-based.	20
	2.3.1.5 Rescoring	20
	2.3.1.6 Docking of fragments	21
	2.3.2 Ligand-based methods	21
	2.3.3 Synergy of in silico methods	22
2.4	Experimental Methods	23
	2.4.1 Surface Plasmon Resonance	23
	2.4.2 Isothermal Titration Calorimetry	23
	2.4.3 Thermal Shift Assay	24
	2.4.4 High Concentration Screening	24
	2.4.5 Mass Spectrometry	25
	2.4.6 Nuclear Magnetic Resonance	25
	2.4.7 X-ray Crystallography	25
2.5	Synergy between in silico and in vitro Methods	27
2.6	Fragment Optimization	27
	2.6.1 Growing	27
	2.6.2 Merging	29
	2.6.3 Linking	30
	2.6.4 SAR by Catalog	31
2.7	Conclusions	31
3	Structure-free optimisation of fragments for the beta-2 adrenergic receptor	36
	3.1 Introduction to GPCRs	36
	3.2 Abstract	41
	3.3 Introduction	42
	3.4 GPCRs- beta receptor	43
	3.5 Chemistry and Biology	44
	3.6 Results and Discussion	46
	3.6.1 Core Optimisation	46
	3.6.2 Docking	49
	3.7 Conclusion	53
	3.8 Acknowledgements	54
	3.9 Supporting Information	55
	3.9.1 Methods	55
	3.9.1.1 Chemistry	55
	General Methods	55
	General Procedure	55
	3.9.1.2 SPR methodology	55
	3.9.1.3 Docking preparation	56
	3.9.2 Results	56
	3.9.2.1 SPR	56
	3.9.2.2 Docking	56
4	PINGUI: Binding-site compatible growing applied to the design of β_2-adrenergic receptor ligands.	62
	4.1 Abstract	62

4.2	Introduction	63
4.3	Methods	65
4.3.1	The β_2 AR binding site	65
4.3.2	Receptor X-ray structures	66
4.3.3	Datasets	67
4.3.3.1	Core fragments	67
4.3.3.2	Building blocks	68
4.3.3.3	Surrogates	68
4.3.3.4	Products	69
4.3.4	Chemical derivatization by reductive amination	69
4.3.5	Docking with SEED	70
4.3.5.1	Sampling and scoring	70
4.3.5.2	Filtering and ranking	70
4.3.6	Docking with DOCK	71
4.3.6.1	Core fragments	71
4.3.6.2	Products	71
4.3.7	Pose minimization: Szybki	71
4.3.8	Workflow of the growing strategy	71
4.3.9	Experimental synthesis	71
4.3.9.1	K011 derivative products	71
4.3.9.2	Z32501319 derivative products	72
4.3.10	Radio ligand displacement assay	73
4.4	Results	73
4.4.1	Creation of the datasets	73
4.4.1.1	Compatible building blocks	73
4.4.1.2	Surrogates	74
4.4.1.3	Products	74
4.4.2	Fragment docking	74
4.4.3	Prediction of the binding mode of the core fragments	74
4.4.3.1	AC molecules	74
4.4.3.2	IAC molecules	75
4.4.4	Structure-based screening of the derivative products	76
4.4.5	Radio ligand displacement assay	76
4.5	Discussion	77
4.5.1	K011FC004	77
4.5.2	K010FC006, K010FC007 and K010FC008	79
4.6	Conclusions	80
4.7	Supporting Information	82
4.7.1	Radio ligand displacement assay	82
4.7.2	Discussion	82
4.7.2.1	K011FC001	82
4.7.2.2	K011FC002	83
4.7.2.3	K011FC006	83
4.7.2.4	K011FC008	84
5	SCUBIDOO	87
5.1	Article	87

5.2	Supporting Information	100
6	Design and identification of novel ligands for the PIM1 kinase	108
6.1	Introduction	108
6.2	Methods	110
6.2.1	The PIM1 binding site	110
6.2.2	Bi-partite product philosophy	110
6.2.3	Growing strategy	111
6.2.4	Receptor preparation	111
6.2.5	Active and decoy sets	111
6.2.6	Libraries preparation	112
6.2.7	Docking	112
6.2.8	Chemical descriptors	112
6.3	Results	112
6.3.1	Evaluation of PIM1 structures by means of enrichment calculations	112
6.3.2	Ligand-based approach	113
6.3.3	Structure-based approach	115
6.3.3.1	Step 1: sample docking.	115
6.3.3.2	Step 2: identification of low-hanging fruits.	115
6.3.3.3	Step 3: deconstruction of the low-hanging fruits.	116
6.3.3.4	Step 4: construction of the building block derivatives.	116
6.3.3.5	Step 5a: docking derivatives and selection of the chemical reaction for growing.	116
6.3.3.6	Step 5b: optimization of the fragment to grow.	117
6.3.3.7	Step 5c: selection of the products for synthesis	118
6.4	Discussion	118
7	Tailored combinatorial synthesis guided by combinatorial growing	124
7.1	Introduction	124
7.2	Methods	125
7.2.1	The beta-2 AR binding site	125
7.2.2	Receptors preparation	126
7.2.3	Libraries preparation	126
7.2.4	Docking	127
7.2.5	Chemical descriptors	127
7.2.6	Datasets	127
7.2.7	Principal Component Analysis (PCA)	127
7.2.8	Synthesis	127
7.2.8.1	1.5 g scale Boc-protection	128
	Equipment required and reaction conditions:	128
	Procedure:	128
	Work up:	128
7.2.8.2	40 mg scale amidification products	128
	Equipment required and reaction condition:	128
	Procedure:	128
	Work up:	128
7.2.8.3	De-Boc	128

Equipment required and reaction condition:	128
Procedure:	129
Work up:	129
7.2.8.4 30 mg scale reductive amination products	129
Equipment required and reaction condition:	129
Procedure:	129
Work up:	129
7.2.9 Selection strategy	129
7.3 Results	130
7.3.1 Binding site partitioning	130
7.3.2 Polar residue fingerprint	130
7.3.3 Identification and selection of the A fragments to grow	131
7.3.3.1 Step 1: sample docking.	132
7.3.3.2 Step 2: deconstruction of the low-hanging fruits.	132
7.3.3.3 Step 3: selection of the organic reactions	133
7.3.3.4 Step 4: construction of the building block derivatives.	133
7.3.3.5 Step 5: docking derivatives.	133
7.3.3.6 Step 6: mini SAR around the A fragments.	133
7.3.4 Identification and selection of the B building blocks to attach	134
7.3.4.1 Amination	135
7.3.4.2 Amide formation	136
7.3.5 Analysis of the generated products	136
7.3.5.1 Principal Component Analysis	136
7.3.5.2 Novelty	137
7.3.6 Synthesis	138
7.4 Discussion	139
7.5 Supporting Information	142
7.5.1 B building blocks for the reductive amination pool	142
7.5.2 B building blocks for the amide pool	145
8 Perspectives	149
8.1 Preamble	149
8.2 Bipartite products and docking disguised fragments	150
8.3 Reverse engineering via molecular obesity	151
8.4 Tailored combinatorial synthesis guided by combinatorial growing	153
8.5 Exploiting the 'loser bracket'	154
8.6 SCUBIDOO future	156
Bibliography	160
Declaration of Authorship	188
Acknowledgements	192

List of Figures

2.1	Typical workflow of the fragment-based approaches in drug discovery . . .	8
2.2	Comparison of MW versus potency for fragment hits, HTS hits, leads and drugs	10
2.3	General binding mode of fragment hits versus HTS hits	10
2.4	Representation of the Lennard-Jones 12-6 function	19
2.5	Fragment based approaches in drug discovery	28
2.6	Linking fragment using growing and merging	30
3.1	GPCRs activation	37
3.2	G-protein-dependent assays	39
3.3	Polar network involved in the activation of the β_2 AR	41
3.4	Decision tree	43
3.5	The initial hits against the β_2 AR	45
3.6	Synthesis of the quinolines	45
3.7	Scaffold Optimisation	45
3.8	The binding affinity and ligand efficiency of analogues with different vectors.	47
3.9	Calculated pKa values of key functional groups.	48
3.10	Modifications to the piperazine ring.	49
3.11	Modifications to the piperazine.	50
3.12	Coloured map of the binding pocket illustrating carazolol	51
3.13	Undecorated 4-piperazine quinoline scaffold	52
3.14	Compound 25 docked into the β_2 AR structure	53
3.15	Data for key scaffold hops.	57
3.16	Binding mode of the undecorated 4-piperazine quinoline scaffold	57
3.17	Compound 10 docked into the β_2 AR structure	58
3.18	Compound 13 docked into the β_2 AR structure.	58
3.19	Compound 12 docked into the β_2 AR structure	58
3.20	Compound 21 docked into the β_2 AR structure	59
4.1	Sliced surface side-view of the β_2 AR binding site	66
4.2	Carazolol and BI167107 in complex with β_2 AR	67
4.3	A fragment hit (Z32501319) chosen for further growing	68
4.4	Illustration of the fragments hits (β_2 AR) chosen for further growing	68
4.5	Illustration of a surrogate derived from an aldehyde	69
4.6	Illustration of a surrogate derived from a ketone	69
4.7	Reductive amination reaction	69
4.8	Illustration of the docking poses filtering	70
4.9	The PINGUI workflow	72

4.10	Prediction of the binding mode of five fragments	75
4.11	Illustration of grown product from PINGUI	76
4.12	Depiction of the eight successfully synthesized products	77
4.13	IC ₅₀ curves from radio ligand displacement assay	78
4.14	Predicted binding mode of the remaining products. (a) K011FC004. (b) K010FC006. (c) K010FC007. (d) K010FC008.	79
4.15	IC ₅₀ curves from radio ligand displacement assay	82
4.16	Prediction of the binding mode of K011FC001	83
4.17	Prediction of the binding mode of K011FC002	84
4.18	Predicted binding mode of the remaining products. (a) K011FC006. (b) K011FC008.	84
6.1	PIM1 crystal structure	110
6.2	PIM1 key residues targeted by our growing strategy.	111
6.3	Comparison of PIM1 crystal structures 2XIY and 3BGP.	114
6.4	Fragment 0FK crystallized in the PIM1 binding site	114
6.5	SCUBIDOO workflow for docking application.	115
6.6	Illustration of a low-hanging fruit identified in a virtual screening of PIM1	116
6.7	Illustration of a derived product of 5175110	119
6.8	Illustration of a derived product of 4012414	119
6.9	Illustration of the derived products in our fragment-growing strategy . . .	120
7.1	Sliced surface view of the binding site of the beta-2 AR	126
7.2	Parallel synthesis illustration	130
7.3	2D depiction of the beta-2 AR binding site with all polar residues.	131
7.4	Polar fingerprint of the beta-2 AR binding site.	131
7.5	Illustration of a low-hanging fruit identified in a virtual screening	132
7.6	Combinatorial growing in the <i>reductive-amination</i> pool	134
7.7	Combinatorial growing in the <i>amide</i> pool	135
7.8	Example of a <i>safe</i> building block via molecular obesity	136
7.9	PCA of the potentially novel ligands for the β_2 AR	137
7.10	Novelty of the <i>amide</i> and the <i>reductive amination</i> pools	138
7.11	Heat map illustrating the purity and amount of the synthesized com- pounds by Taros.	139
8.1	<i>Unspecific</i> building blocks	152
8.2	Fragment falling to the loser bracket	155

List of Tables

2.1	Summary of the different possible scenarii when starting a ligand discovery project.	14
4.1	Summary of the activity of the core fragments and their respective derivative products.	76
6.1	Enrichment of a divers set of PIM1 structures.	113

Abbreviations

ADMET	A bsorption D istribution M etabolism E xcretion T oxicity
ADP	A denosine D i P hosphate
ATP	A denosine T ri P hosphate
β_2 AR	B eta-2 A drenergic R eceptor
cAMP	cyclic A denosine M ono P hosphate
CNS	C entral N ervous S ystem
CYP	C Ytochrome P 450
ECFP	E xtended- C onnectivity F inger P rints
ECL	E xtra C ellular L oop
ER	E ndoplasmic R eticulum
FBDD	F ragment B ased D rug D iscovery
FRET	F luorescence R esonance E nergy T ransfer
GA	G enetic A lgorithm
GDP	G uanosine D i P hosphate
GPCRs	G P rotein- C oupled R eceptors
GTP	G uanosine T ri P hosphate
HBA	H B ond A cceptor
HBD	H B ond D onor
HCS	H igh C oncentration S creening
HTS	H igh T hroughput S creening
ICL	I ntra C ellular L oop
IP	I nositol T riphosphate
ITC	I sothermal T itration C alorimetry
LE	L igand E fficiency
MC	M onte C arlo

MD	M olecular D ynamics
MLR	M ultiple L inear R egression
MS	M ass S pectrometry
MST	M icro S cale T hermophoresis
MW	M olecular W eight
NCE	N ovel C hemical E ntities
NMR	N uclear M agnetic R esonance
PAINS	P an- A ssay I Nterference compound S
PDB	P rotein D ata B ank
PIM	P roviral I ntegration M aloney
PINGUI	P ython I n silico de N ovo G rowing U tilities
PLC	P hospholi P ase C
PMF	P otential of M ean F orce
PSA	P olar S urface A rea
QSAR	Q uantitative S tructure- A ctivity R elationship
QSPR	Q uantitative S tructure- P roperty R elationship
RDA	R epporter- D isplacement A ssay
RMSD	R oot- M ean- S quare D eviation
RO3	R ule O f T hree
RO5	R ule O f F ive
SAR	S tructure A ctivity R elationship
SAS	S olvent A ccessible S urface
SBP	S econdary B inding P ocket
SCUBIDOO	S creenable C hemical U niverse B ased on I ntuitive D ata O rganizati O n
SEED	S olvation E nergy for E xhaustive D ocking
SPR	S urface P lasmon R esonance
STD-NMR	S aturation- T ransfer D ifference N uclear M agnetic R esonance
TBP	T ertiary B inding P ocket
TSA	T hermal S hift A ssay
vdW	v an d er W aals

Dedicated to my beloved family...

"We've always defined ourselves by the ability to overcome the impossible. And we count these moments. These moments when we dare to aim higher, to break barriers, to reach for the stars, to make the unknown known. We count these moments as our proudest achievements. But we lost all that. Or perhaps we've just forgotten that we are still pioneers. And we've barely begun. And that our greatest accomplishments cannot be behind us, because our destiny lies above us."

Cooper, *Interstellar* (2014)

Chapter 1

Introduction

What are fragments? How can one utilize them in a drug discovery endeavor? How can *in silico* approaches assist us with such efforts? How can one improve the reliability of fragment docking?

Only a few questions, and yet so much that needs answering to... I pursue to deliver responses to those questions in the following thesis.

This first chapter aims to walk you through the story of my PhD and hopefully give you a taste of it within a few minutes. Every project, and thus chapter, relies on the previous ones, making the results increasingly exciting (at least from my point of view).

The second chapter is the cutting edge of fragment-based drug discovery (FBDD). Thus, I will highlight the limits of the current high-throughput screening technique (HTS), and show how fragment-based efforts can go beyond those limits in order to provide novel and potent chemical entities that could act as reliable starting points in any drug discovery endeavors. I will cover experimental and *in silico* screening techniques that allow one to identify fragments and discuss the different fragment extension strategies. This chapter will be closed with the ongoing challenges, as well as the perspectives in the FBDD field.

Chapter 3 is made up of two part. First, I will introduce the GPCRs, which are critical biological targets in the field of pharmaceutical research. The main target in my thesis was the β_2 AR, three chapters and thus projects are devoted to it. The second part consists of a submitted manuscript where I had my first opportunity to work with experimental collaborators. In this work, novel nanomolar ligands were synthesized and tested and I used docking as a 'third' technique in order to try to rationalize the experimental results (SAR). The main message will be that docking cannot explain everything and clearly shows limits when it comes to compare micromolar with nanomolar compounds in a SAR study.

Chapter 4 was my main project: Growing a fragment with favorable ligand efficiency (LE) into more efficient ligands for the β_2 AR. This was my longest project (i.e. about three years) and probably the one I learned the most from. Many reasons explain such a duration, but the most significant one is probably my initial lack of pragmatism. Most of my original work won't appear in this manuscript, but can be briefly summarized. *De novo* tools assisting fragment growing are great when it comes to generate millions or even billions of virtual molecules, but they lack synthetic accessibility filters. It is one thing to create a molecule using a computer, it is another thing to translate it to real existing molecules (i.e. wet chemistry). My initial workflow took into account synthetic accessibility, but only in the last stage. This simply meant that I had to spend hours on retrosynthetic studies in order to provide a synthetic route (if any) for a single compound. This was a giant time sink, and such procedure could not be applied to a large set of molecules. To make a long story short, at some point, in any synthesis project, you will have to sit down with your chemist and discuss about your molecules. So why not make this step easier, and thus take synthetic feasibility into account right from the start? I then found the work of Hartenfeller et al. [1, 2], where the authors compiled a set of 58 organic reactions, based on the most popular reactions used in the pharmaceutical field. At the moment, this is probably the most underestimated paper I ever read. Maybe I overemphasize it because it brought an obvious solution to my problem, but I have to say it was odd to me that this set was not applied in many projects and only cited a few times. Anyway... this set allowed me to identify a compatible reaction for my fragment (*reductive amination*) and we were able to suggest reliable extensions to our chemist partners. All suggested products were successfully synthesized (8 / 8) and half of the products exhibited a better affinity than the original fragment. It was a successful project at two levels: synthesis and experimental assays. In my humble opinion, the most critical one was the successful synthesis. Indeed, being able to suggest novel molecules to chemists and see them endorse those after a quick glance, somehow brings satisfaction. I then decided to make most of the tools I developed in my workflow available online (PINGUI), so computational and experimental chemists alike could freely utilize them. Drug discovery is built on three main pillars, namely computational techniques, chemistry and biology. I like to think that PINGUI allows one to bring closer the computational techniques and chemistry, and can be beneficial in both fields. Indeed, computational chemists will have insights into organic reactions early on, thus increasing their chance to come up with molecules that are likely to be synthesizable. This will also make the synergy with the chemist more powerful, because their feedback will be taken into account early on in the process. As for the chemists, PINGUI could be used as a creative tool to bring more ideas on how to decorate chemical scaffolds.

Chapter 5 was the consequent step following the PINGUI workflow. Since *reductive amination* worked well in this project, we simply asked ourselves: 'why not expand the concept?'. Indeed, we only used one reaction out of 58, so let's take advantage of the remaining ones. To do so, we decided to start from a small library of available building blocks (about 8'000) and exhaustively combined them with each other using the 58 reactions. This procedure gave birth to SCUBIDOO, which currently holds 21 M virtual products. Three representative samples of this database were created in order to give the user entry points to this new virtual chemical space. Retrospective analysis showed that known ligands were contained in the database and that the predicted synthetic route was also correct. Only two weeks after its publication, SCUBIDOO was recommended in F1000 and associated with the 'interesting hypothesis' and 'technical advance' flags. One of the reviewers described it as 'a new welcome resource in the field'.

Chapter 6 was the first 'real' application of SCUBIDOO and the biological target was the kinase PIM1. This project initially had three main goals. First, showcase how to use SCUBIDOO and provide to the scientific community a detailed guide of all possible scenarios. Second, this project should be a proof of concept highlighting the synthesizability of the products. Third, identify fragments with favorable LE that we can quickly optimize. Three fragments were identified for further optimization, each using a different strategy. One was found using ligand-based approaches (similarity search), one was identified using structure-based techniques (docking) and the last one was discovered when trying to improve the hit from docking (analog search). Each fragment was associated to a different chemical reaction. As of right now, no derivatives were synthesized, but one of the initial fragments was tested in a thermal shift assay (TSA) and showed a positive shift of + 1.8°C, which can be taken as a first experimental hint that this fragment could bind to PIM1. The first attempt to crystallize this complex was successful, which could be taken as a second experimental hint. Crystals will be solved in the near future in order to hopefully validate the predicted binding mode of this fragment. Even though, so far, no derivatives were synthesized, this project offers a silver lining: a fragment hit was identified through docking by making use of the 'bipartite philosophy' (i.e. a product is the assembly of two building blocks). Indeed, a virtual product was identified with one half making compelling interactions while the other half did not. Deconstruction of this product yielded the identification of the 'promising' fragment, which was then selected for optimization. While this project is still ongoing, I hope it will describe new *in silico* guidelines to identify fragments paired with favorable LE.

Chapter 7 was the second application of SCUBIDOO, but the first one at a large scale. This project represents quite a leap forward in terms of means, in comparison to chapter 6. We had the opportunity to work with an industry partner specialized in custom synthesis (Taros, Dortmund). Taros has automated robots for synthesis and can process

24 reactions in parallel. Hundreds of compounds can be synthesized and purified within a day. As mentioned in chapter 5, we suggested that SCUBIDOO could be integrated into such an automated synthetic workflow as an idea generator. Thus, this project was a fantastic follow-up and we decided to apply it to the β_2 AR, because we have a lot of experience with this target. The biggest challenge of this project was to define an initial strategy, because none such studies were reported in the literature. In the end, it could be summarized as 'tailored combinatorial synthesis guided by combinatorial growing'. It could have been called the other way around, but in our case the synthetic constraints guided the design of the compounds, hence synthesis was the first concern and then we had to adapt in order to suggest potential ligands for the β_2 AR. Three main goals with increasing difficulty were initially defined. First, validate the synthetic feasibility of SCUBIDOO products. Second, identify ligands and fragments with favorable LE. Last, design high affinity ligands. 127 out of 240 products were synthesized (53%) highlighting the strong potential of SCUBIDOO to design novel compounds in an very short time frame. Experimental assays will take place in the following weeks, and we aspire to have the first results before the fall of 2016.

The last chapter can be seen as an extended discussion of chapters 5, 6 and 7. Those three projects were probably the most exciting part of my thesis. Indeed, I developed two scientific tools, then applied them to concrete scientific efforts, then learned from these projects and made some unexpected discoveries. From this, new ideas logically came out in order to improve SCUBIDOO. This quickly became an iterative process which evolved into a virtuous loop. I wanted the last chapter to reflect that. Most of the discoveries detailed here are still speculative, but should be experimentally validated (or not) in the near future. This chapter also aims at describing SCUBIDOO future applications and possible improvements. After all, SCUBIDOO is still young and showed a lot of potential, so I like to think that I am only writing the beginning of a story.

"Success consists of going from failure to failure without loss of enthusiasm."

Winston Churchill

Chapter 2

Fragment Based Drug Discovery

2.1 Introduction

Efficacy and selectivity are two of the main challenges that the pharmaceutical research field is facing with the need to design novel chemical entities (NCE) which are effective drugs for many diseases, combined with as few side effects as possible [3]. In parallel, the cost to deliver a drug to the market is now estimated at more than \$2.5bn. Most alarming about this observation is the fact that this price has more than doubled in the past 10 years [4]. Two reasons may explain the ever so increasing cost: the time required to develop a drug, which is estimated between 8 and 15 years, and the high risk of failure in both pre-clinical development and clinical trials. The average clinical approval success is estimated to be below 10%. This success rate will depend on the therapeutic area, ranging from 8% for Central Nervous System (CNS) drugs to 24% for systemic anti-infective drugs [5]. The further a drug goes through the drug discovery pipeline, the more costly a failure will be, since the invested means increase over time.

One of the main reasons that could explain this high attrition rate resides in High Throughput Screening (HTS) techniques, which have reached their limits. HTS consists of screening thousands of drug-sized molecules against a therapeutic target of choice in an automated fashion. Pharmaceutical companies have been using this approach for decades. The main advantage of this method is that it allows one to deal with a large number of existing drug-like molecules within a short time frame, with up to 100'000 compounds per day. However, screening 100'000 compounds is costly and not within reach of academic institutions. Moreover, the screened molecules represent only a small fraction of chemical space. This is due to the fact that those libraries contain a lot of molecules which were optimized for historical targets, thus leading to low chemical diversity [6]. Furthermore, most of those molecules are in principle drug-like, rooted in

Lipinski's rule of five (RO5) [7, 8]. This rule describes the likelihood of oral bioavailability for any given compound, based on simple chemical descriptors: Molecular weight (MW) ≤ 500 Da, $\log P \leq 5$, number of H-bond donors (HBD) ≤ 5 and number of H-bond acceptors (HBA) ≤ 10 . A HBD is defined by an hydrogen atom attached to a heteroatom (nitrogen or oxygen), while an HBA is defined by an heteroatom with at least one lone pair available (sulfur is also considered as acceptor, but weaker than N or O). Once a hit (i.e. a molecule exhibiting biological activity) is identified, optimization follows, aiming at finding leads (i.e. a hit with an improved biological activity) which could become potential drugs. But even though such molecules are in principle drug-like, they will have likely unfavorable absorption, distribution, metabolism, excretion and toxicity (ADMET) properties [9]. The problem with such hits is that they already feature a high molecular weight, making the process of optimization harder since they are already close to the Lipinski limit. Furthermore, in order to increase potency, medicinal chemists tend to increase the lipophilic character of compounds. This problem is known as molecular obesity [3], which leads to the failure of such compounds in the later stages of drug development, often due to a lack of selectivity.

But why focus on big molecules and try to improve them with high attrition rates, when one could start from smaller, yet still potent, molecules and try to extend them in order to improve potency? This approach is known as fragment-based drug discovery (FBDD) as illustrated in figure 2.1.

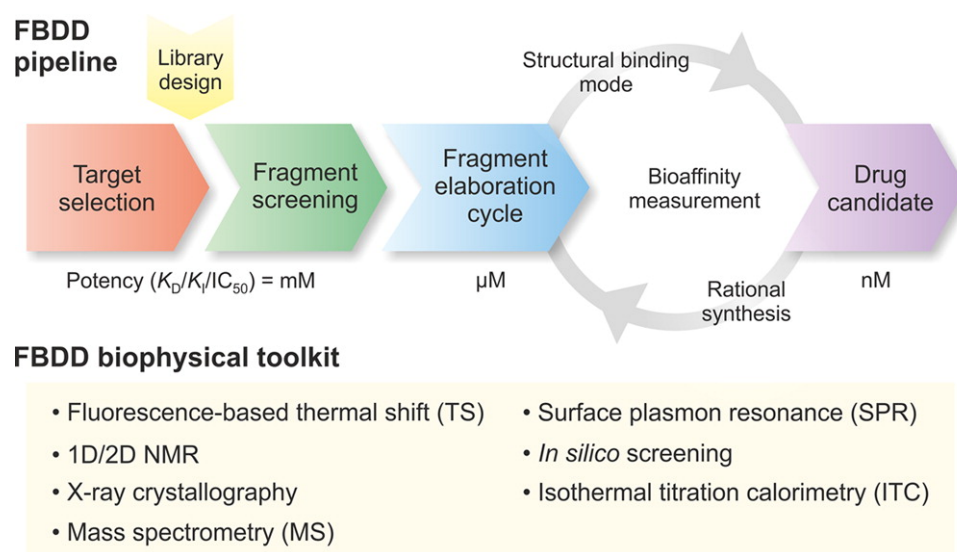


FIGURE 2.1: Typical workflow of the fragment-based approaches in drug discovery. Reprinted with permission from Duncan E. Scott, Anthony G. Coyne, Sean A. Hudson, and Chris Abell. Fragment-based approaches in drug discovery and chemical biology. *Biochemistry*, 51(25): 4990–5003, 2012. Copyright 2016 American Chemical Society.

In the past two decades, FBDD has continuously gained popularity in drug discovery efforts and has become a dominant tool in order to explore novel chemical entities that

might act as bioactive modulators [6, 10]. Shuker *et al.* are among the pioneers in this field with their *SAR by NMR* approach [11], where they identified nanomolar affinity ligands for the FK506 protein by linking together two small ligands (fragments) with micromolar affinity each.

But what exactly is a fragment? By its very own nature, a fragment is a small entity part of a bigger puzzle. Applying this concept to molecules simply results in small molecules prone to be extended. Same as for drug-like compounds, there soon appeared a discussion on which chemical properties should define a fragment. Congreve *et al.* came up with a simple rule of three (RO3) [12] (Molecular weight $\leq 300Da$, $\log P \leq 3$, H-bond donors ≤ 3 and H-bond acceptors ≤ 3), directly inherited from the RO5.

Fragments offer many advantages compared to most compounds in typical HTS collections. Due to their small size, fragments are more amenable to make specific interactions compared to bigger molecules (figure 2.3), simply because of an increased likelihood that all their (few) chemical features are complemented. Even though fragments usually bind within the millimolar to micromolar range, the ligand efficiency (LE) is remarkably high (figure 2.2). LE is defined as the contribution to free energy of binding by each heavy atom [13–15] as defined in equation 2.1:

$$LE = \frac{-2.303RT}{HA} \cdot \log K_d \quad (2.1)$$

where R is the ideal gas constant, T is the temperature in Kelvin, HA is the number of heavy atoms and K_d is the equilibrium dissociation constant (which is inversely related to affinity). Fragments do not only offer optimized interactions, they also contain fewer interfering moieties, making them more reliable in screening campaigns [6]. Moreover, it is easier to cover chemical space with fragment-sized molecules: screening a diverse set of 1’000 fragments is equivalent to probing the chemical space of 1’000’000 drug-like molecules, yielding a higher hit rate [6, 16–18]. Thus, fragment screening is cheaper than HTS, making it more affordable and also a good fit for academic institutions. An example of the coverage of chemical space with fragments is the creation of SCUBIDOO (chapter 5), where 7’805 building blocks were combined yielding 21M virtual products.

Finally, FBDD could be summarized by “*start small and stay small*”. Optimizing incrementally very efficient fragments keeps molecular obesity at bay and yields leads offering better ADMET profiles. Therefore, leads coming from an FBDD campaign are less likely to fail compared to hits retrieved from HTS, since they offer better control over the design process [19]. This statement will require confirmation over the next couple of years, but the first tendency already shows encouraging results. An example is Vemurafenib,

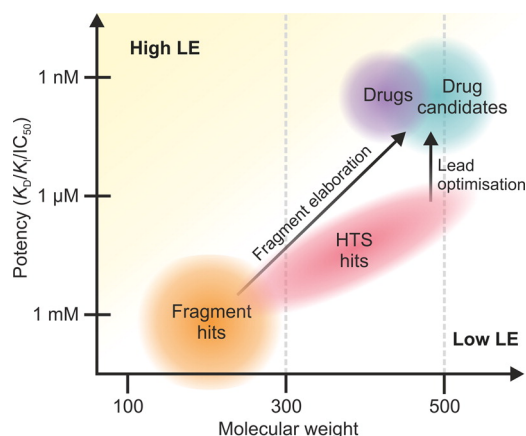


FIGURE 2.2: Comparison of MW versus potency for fragment hits, HTS hits, leads and drugs. Reprinted with permission from Duncan E. Scott, Anthony G. Coyne, Sean A. Hudson, and Chris Abell. Fragment-based approaches in drug discovery and chemical biology. *Biochemistry*, 51(25): 4990–5003, 2012. Copyright 2016 American Chemical Society.

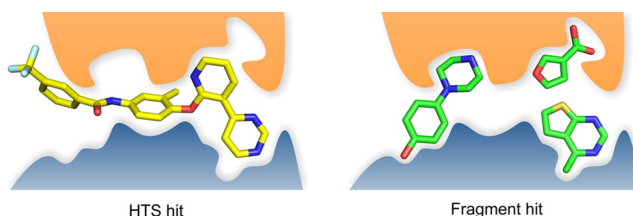


FIGURE 2.3: Typical binding mode of fragment hits versus HTS hits. Reprinted with permission from Duncan E. Scott, Anthony G. Coyne, Sean A. Hudson, and Chris Abell. Fragment-based approaches in drug discovery and chemical biology. *Biochemistry*, 51(25): 4990–5003, 2012. Copyright 2016 American Chemical Society.

the first FDA-approved drug [20] originating from a fragment-based screening. Vemurafenib was approved for BRAF-mutated metastatic melanoma, was marketed in 2011 and it took only 6 years to bring this drug to the market.

While FBDD looks extremely appealing, some aspects remain challenging, especially in the experimental field. Since fragments bind with low affinities, more sensitive detection methods are required as well as screening the fragments at high concentrations. This raises the concern of solubility: in order to be able to screen at high concentrations, each fragment ought to be soluble enough. Also, at high concentration, compounds may form aggregates which can pretend unspecific enzyme inhibition and yield false positive hits [21, 22]. This is a known artifact from experimental screening [23]. Another limit is that experimental assays need to be sensitive enough to detect fragments and thus often require a large amount of purified proteins (from 10 mg to 1g), restraining the number of targets that can be screened [6].

In this chapter, I will cover the vital steps that are needed for any fragment-based ligand discovery effort. I will first describe how to create a fragment library. Next, I will review

the methods in both the *in silico* and experimental fields, which are used to identify fragments that can serve as starting points. I will then highlight the synergy that is provided when using both methods in parallel. Then, once a fragment hit is identified, I will describe the different ways to optimize it. Finally, I will conclude with the current challenges that FBDD is facing and how my work might help in the current fragment elaboration/optimization cycle.

2.2 Fragment libraries

Prior to any fragment screening is the conception of the fragment library. Fragments can be filtered according to several chemical properties which can be predicted or calculated by computer approaches. Fragment libraries have two main applications: screening (experimental or virtual) and exploration of chemical space using combinatorial techniques.

2.2.1 Important properties for experimental screening

2.2.1.1 Library size

The size of a fragment library depends on its application, and more precisely on which kind of experimental assays will follow up. Recently, Keseru et al. [24] suggested that a reasonable number could be between 500 and 3000 fragments. For instance, a smaller library will be more appropriate for lower throughput screening techniques such as X-ray crystallography. An illustration of such small library, is the in-house fragment library of the Klebe group which currently holds 361 fragments and is extensively used for X-ray screening [25–27]. On the other hand, bigger libraries are more appropriate for higher throughput methods (e.g. SPR).

2.2.1.2 Chemical diversity

Creating a fragment libraries of chemically diverse compounds offers three main advantages:

- Ensure to cover as much drug-like chemical space as possible.
- Ensure diversity among the chemical features. This will allows one to explore different binding modes but also ensure the application of different organic reactions, which will facilitate the extension procedure.

- Prevent the presence of close analogues, which could yield duplicates in binding modes.

To do so, representative subsets can be extracted from existing libraries using dissimilarity techniques or clustering [28], based on molecular fingerprints such as MACCS keys [29] or ECFP [30]. The MACCS keys are particularly well suited for this task since they encode the presence of 166 predefined chemical features. More sophisticated approaches could be used based on more complex chemical descriptors (2D or even 3D) or different algorithms, such as stratified balanced sampling [31–34], which will be fully described in chapter 5.

2.2.1.3 Solubility

As mentioned above, fragments should exhibit high solubility (usually well above 1 mM) in order to be screened at high concentrations [18, 35–37]. When solubility has not been experimentally determined, one can predict it using *in silico* models. While prediction of solubility remains challenging [7, 38–40], the accuracy of the models can be improved by multimodel protocols, as I highlighted in a previous study [41].

2.2.1.4 Reactive functional groups

Since fragments are prone to be extended, establishing the presence of reactive chemical features (i.e. functional groups that can be utilized to attach a chemical group to) is important for later optimization. I implemented a tool in that regard (PINGUI chapter 4), aiming at suggesting which chemical reaction could be applied to a fragment in a growing strategy context. Once compatible reactions have been identified, this tool creates libraries of derivatives of the initial fragment. Therefore, the generated products are optimized towards high likelihood of synthetic feasibility.

2.2.1.5 Rule of three compatibility

From a strict point of view, fragments could be defined using the RO3 and filtering the library according to those chemical properties could also be done. However, the rule of three might be considered a bit too rigorous and might discard potentially fragment-sized binders. An example of this was the creation of a non-rule-of-three compatible fragment library by the Klebe group [27]. The fragment library contains 361 entities and was designed in a way that fragments are likely to be successful for crystallographic screening and follow-up chemistry (i.e growing or merging). The biggest violation of

the rule of three concerned the number of H-bond acceptors and rotatable bonds. Both were allowed up to 7, and 223 fragments did not agree with the RO3. Yet, the library was screened against endothiapepsin using a fluorescence-based competition assay and yielded 55 hits, which is remarkably high. Those 55 entities were then soaked into native endothiapepsin (ETP) crystals and yielded 11 crystal structures [27]. More recently, the same group exhaustively soaked all the 361 fragments individually into ETP and they were able to crystallize 71 structures (20% hit rate !). Those results suggest that the composition of the initial library was efficiently done and, most importantly, highlight that the RO3 is better be used as a general guideline rather than a strict rule.

2.2.1.6 Specialized application

Specific considerations in library design might have to be taken into account. For instance, ^{19}F -NMR screening is an experimental technique that require the presence of fluorine containing fragments. This experimental method is relatively new and start to be implemented in FBDD efforts quite successfully [42–45]. Chemical suppliers adapted quickly to that new demand and often offer fluorinated fragments libraries ready to be screened. For instance, Enamine proposes a RO3 library of 3'000 ^{19}F containing fragments.

2.2.2 Important properties for virtual screening

In silico approaches can be utilized to screen more fragments more quickly than experimental techniques and can also screen virtual fragments. However, one always need the experimental techniques to ultimately validate a fragment hit. For these reasons, the properties required for virtual screening libraries inherit from the experimental properties libraries. The only difference would be a lower emphasis on a preselection according to high chemical diversity, since one can easily apply an exhaustive approach in virtual screening. Indeed, screening millions of fragments can be done within a few hours on a computer cluster.

Additional filters that deal with toxic features or pan assay interference compounds (PAINS) [46] are also important when preparing a library of fragments. However, such filters are better implemented downstream, since toxic or reactive features may vanish or appear when a product is generated (i.e. a grown fragment).

There are no commonly agreed chemical rules on how to design a fragment library, but a recent guideline was suggested by Keseru et al. [24] highlighting that it requires the consideration of multiple parameters and thus one often has to find compromises.

The design will mainly depend on the context as well as the subsequent application. Factors such as purity, stability and storage conditions ought to be considered as well. One also has to keep the big picture in mind: if one obtains a fragment hit, one will very likely extend it. Therefore, most of the attention has to be focused on possible synthetic extensions: where does one grow a fragment from ; how (i.e. reaction) ; is there is enough space ? Those important questions are illustrated and answered with the PINGUI workflow (chapter 4).

2.3 In Silico Approaches

Several scenarios are possible when starting a ligand discovery project (Table 2.1). In all cases, *in silico* approaches provide valuable support all along the way.

Scenario	Crystal structure ?	Crystallized ligand ?	Known active ?
1	✗	✗	✗
2	✗	✗	✓
3	✓	✗	✗
4	✓	✗	✓
5	✓	✓	✓

TABLE 2.1: Summary of the different possible scenarii when starting a ligand discovery project.

For scenario 1 and 2, when the crystal structure of the target has not yet been solved, alternative approaches such as homology modeling are available [47]. This approach allows one to create the three-dimensional structure of a protein based on the knowledge of resolved structures with similar sequence. However, if it is impossible to construct a reliable model of a protein target, we can still hope to identify bioactive fragments using ligand-based methods.

Once the the structure of the target protein is known, or its modeling is reliable enough, one can start structure-based approaches in order to identify potentially active compounds. Robustness of the receptor or its model can be evaluated using computational approaches such as the calculation of the enrichment of known ligands over decoys. This procedure aims at estimating how well a receptor is able to discriminate known active molecules against inactive ones (decoys).

In the context of fragment-based drug discovery projects, if the fragment of interest has been solved in complex with its target, this represents the best-case scenario. Indeed, by knowing the precise binding mode of a fragment combined with an understanding of the binding site, one can quickly come up with extension suggestions.

When no experimental data are available for the fragment interactions, its binding mode within an active site can be predicted using *in silico* methods, most notably docking.

2.3.1 Structure-based methods

Docking aims at predicting the binding mode of a ligand within the binding site of a protein, as well as the free energy of binding of the resulting complex. The latter still remains extremely challenging [48]. The process of docking can be divided into two major subprocesses: the *sampling* phase (or *posing*), where a given conformation is placed into a binding site and the *scoring* phase, where the free energy of binding of each pose has to be rapidly estimated. Prior to sampling, the binding site has to be defined and different conformations ought to be generated for the ligand.

2.3.1.1 Defining the binding site

If the crystal structure of the target is available with a ligand, the definition of the binding site is straightforward. The residues surrounding the ligand (usually within 5 Å) are used to define the binding site. This can be set by the user optionally. This procedure can be automatized with tools such as DAIM [49].

If no information about the ligand is available, computational approaches can help to predict the binding site. There is a plethora of software tools available, most of which are freely accessible online, among them Q-site finder [50], Ligsite [51], F-pocket [52, 53] or DogSiteScorer [54–56].

2.3.1.2 Exploring the conformational space of the ligand

Two classes of docking programs need to be distinguished when dealing with conformer generation: Flexible ligand and rigid ligand. Programs for flexible docking usually generate the conformers on-the-fly, while rigid docking programs need a second software tool to generate them. Two main classes of algorithms are employed to generate ligand conformers: Exhaustive and stochastic search algorithms.

Exhaustive search. This approach aims at exploring all the possible degrees of freedom of the ligand. OMEGA [57] is an example of this method. OMEGA relies on torsion and ring libraries to identify rotors and flexible rings [58]. A filter is applied to remove internal clashes as well as energetically unfavorable conformers. Several custom settings can be modified such as the number of generated conformers, the minimum

root-mean-square deviation (RMSD) between each generated conformer or the strain energy. Other approaches based on rotamer libraries consist of methods that assign the most likely angle value to rotors based on atom types [59–63].

Stochastic search. Rotors are the key to exploring conformational space. Another way to generate conformers is to randomly assign values to the rotors. This can be done using evolutionary algorithms [64–71] or Monte Carlo simulations (MC) [72–74].

In a molecular context, genetic algorithms (GA) are an often used evolutionary approach [75]. They apply the principle of biological competition as criterion to extract individuals from a population. The rotors are encoded as genes on a chromosome and are then randomly varied. Thus, chromosomes yield possible conformers (solutions) that are evaluated using a fitness function. Only the best solutions are selected for improvement using crossover (exchange) or mutation operations in order to create the next generation. This process reiterates for each new generation.

Monte Carlo simulations start with an initial conformation and aim at improving it by random perturbations of the rotors in an iterative fashion. Perturbations are accepted or rejected based on a Boltzmann probability [76]. MC simulations are rather slow, which make them not really suited for large screening campaigns [60].

2.3.1.3 Orientational sampling

Sampling consists of determining whether the position and the orientation of a given conformer will fit the binding site of a protein (the latter is most of the time considered rigid). This procedure is rather fuzzy and will result in many poses that will be filtered later on during the scoring phase. Several approaches exist to place the conformer within a binding site, and most of them rely on trying to satisfy and improve the interactions of key chemical features.

Systematic and random searches. As for the conformational space of the ligand, the translational degrees of freedom defining the position and orientation of the ligand can be explored exhaustively or randomly.

SEED [77, 78] (Solvation Energy for Exhaustive Docking) is classified as an exhaustive search method for fragment docking. SEED uses polar and apolar vectors to describe the fragment and the binding site. Polar vectors are defined as originating from a polar atom. The length and the orientation of the vectors are then based on all the favorable angles and distances to establish an H-bond based on the involved atom types. Vectors

that point towards occupied regions of space (i.e. receptor) are discarded. The sampling phase consists of matching the polar vectors of the fragment with the ones from the binding site, so that the distance between the H-bond donor and the H-bond acceptor is favorable with respect to the atom types. The fragment is then rotated around the H-bond axis and the user has control over the number of rotations around each axis. Apolar vectors are defined in a two-steps procedure. First, points are distributed uniformly on the solvent-accessible surface (SAS) of the receptor binding site and the ligand. Secondly, a low dielectric sphere (probe) is run over the aforementioned points in order to evaluate the desolvation energy, and the van der Waals (vdW) interaction with the receptor. Only the best points according to the two energetic terms are kept. Apolar vectors for the fragment and the receptor are then defined by joining each point on the SAS with its corresponding atom center. The sampling consists of matching apolar vectors of the fragment with the ones of the binding site, so their van der Waals distance is optimal. As in the case of polar vectors, the fragment is then rotated around the axis defined by the fragment atom and its receptor counterpart.

GAs can also be used to perform random searches in this case. In this scenario, the degrees of freedom defining the position and orientation of the ligand can be encoded within the chromosomes [64, 79].

Shape complementarity methods. FRED [58, 80–82] belongs also to the systematic search approaches, but it complements it with a shape filter. In a first stage, all rotations and translations of the ligand are exhaustively sampled. During a second stage, a so-called 'negative image' of the binding site is created. To do so, molecular probes that represent common chemical features of known drugs (i.e. pharmacophores) are placed within the binding site and scored according to the Gaussian Shape Scoring Function [80]. The very best poses are converted into density fields and merged together in order to form the final shape of the 'negative image'. Thus, when comparing the ligand pose to the 'negative' image, we expect high values of the potential field where ligand atoms make favorable contacts with the receptor as well as very little steric overlap.

Point complementarity methods. DOCK 3.x relies on the representation of spheres to encode potential atom placements [62, 83, 84]. The binding site is filled with a set of spheres that smoothly interact with the surface of the receptor. The ligand representation is also based on a set of spheres that allows to depict its shape. Thus, if the set of ligand spheres fits the set of receptor spheres, one could assume that the protein-ligand interactions will be favorable. The pairing of the protein and ligand spheres is possible if the internal distances of all ligand spheres match the internal distances of the receptor spheres (within a certain margin of error).

Incremental methods. Incremental approaches [61, 62, 85] are distinct from all the aforementioned ones, since the ligand, and thus its 3D conformer, is generated on-the-fly. For instance, in FlexX[63, 86] the ligand to dock is first split into several fragments. A first fragment, usually the one that can be assumed to make the most compelling interactions, is placed within the binding site and meticulously optimized. Each remaining fragment is attached in an iterative process. At each step, the ligand is exhaustively optimized and only the best solutions are kept.

2.3.1.4 Scoring

Once poses have been sampled and the obvious clashes removed, one needs to evaluate rapidly which ones are the most likely to bind favorably. Scoring functions were developed for that purpose and aim at giving a rough measure of the fit of the ligand pose within the binding site. Typically three classes of scoring functions are used: force-field-based, empirical and knowledge-based scoring functions.

Force-field-based. A force-field is a set of parametrized potential functions describing the mechanics of a molecular system. Force-fields are usually applied in the docking context to estimate the energy of binding between the ligand and the receptor along with the internal energy of the ligand. Usually, ligand-protein non-bonded interactions and internal ligand energy are described using van-der-Waals and electrostatic energy terms.

The van-der-Waals energy component is often modeled with a Lennard-Jones 12-6 potential function, as described in equation 2.2:

$$E_{vdW}(r) = \sum_{i=1}^N \sum_{j=i+1}^N 4\epsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (2.2)$$

In equation 2.2, ϵ represents the well depth of the potential as illustrated in figure 2.4 and σ is the contact distance between the atoms i and j . The Lennard-Jones 12-6 potential function is considered 'hard' and will severely penalize close contacts between ligand and receptor atoms. Some programs such as FRED [58], prefer a 'softer' potential defined by an 8-4 term. Softer potentials were shown to be less accurate than 'harder' ones in retrospective docking studies and often generate poses with steric clashes. However, for hydrophobic binding sites, FRED was found to produce more accurate results because hydrophobic interactions outweigh electrostatic and polar interactions [75]. We illustrate this precise scenario in chapter 6, where we identified a fragment hit using FRED.

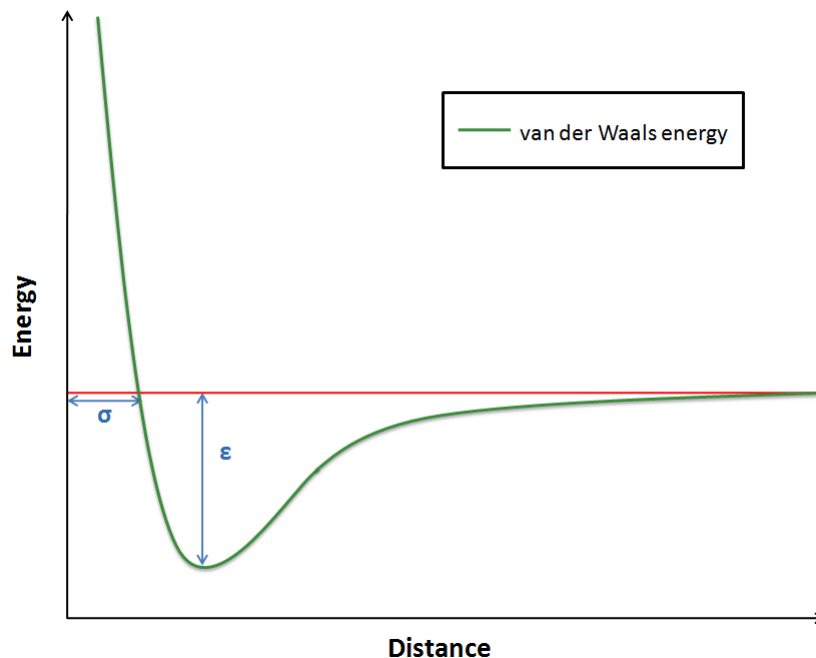


FIGURE 2.4: Representation of the Lennard-Jones 12-6 function. The $\exp(12)$ term of the equation represents small-distance repulsion (the part of the curve above the red line), while the $\exp(6)$ term represents long-range attraction between two atoms (the part of the curve under the red line).

Electrostatic potential energy is modeled through a pairwise summation of Coulombic interactions, as described in equation 2.3:

$$E_{coul}(r) = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij} \epsilon_m} \quad (2.3)$$

In equation 2.3, N_A and N_B represent the number of atoms in molecules A and B respectively, $q_i q_j$ the partial charge on each atom, r_{ij} the distance between the point charges, ϵ_0 is the dielectric constant of the vacuum and ϵ_m is the dielectric constant of the medium.

Additional terms accounting for conformational entropy and desolvation of the ligand and the receptor are taken into account for the final score. However, their prediction is often dilemma since one had to find a compromise between accuracy and speed. Since docking is oriented towards virtual screening of millions of molecules, the speed is often the priority and one had to rely on models with poor predictive power.

Since most of the docking programs are using rigid receptors, there is no internal protein energy to compute. This is one of the main limitations of most docking approaches that can be partially answered either by giving some flexibility to the residues of the binding site [87], or by treating the protein as an ensemble of protein conformations extracted from molecular dynamics (MD) [88].

Empirical. These scoring functions intent to reproduce experimental values of binding free energies. The rationale behind empirical scoring functions is that the binding free energy can be decomposed into a sum of uncorrelated variables [63, 89–93]. The coefficients associated with each variable are computed using regression analysis techniques, such as multiple linear regression (MLR). For instance, Rognan *et al.* developed FRESNO [91] to predict the binding free energy of peptides to class I major histocompatibility (MHC) proteins. This model is based on several weighted ($\alpha, \beta, \gamma, \delta, \epsilon$) contributions: a constant (K), H-bond (HB), lipophilic interactions ($LIPO$), torsional entropy (ROT), buried polar surface (BP), solvation and desolvation effects ($DESOLV$) as illustrated in equation 2.4:

$$\Delta G_{bind} = K + \alpha(HB) + \beta(LIPO) + \gamma(ROT) + \delta(BP) + \epsilon(DESOLV) \quad (2.4)$$

However, empirical scoring functions are strongly dependent on the training set used for fitting the values, resulting often in one or more scoring functions developed for each individual protein target.

Knowledge-based. These scoring functions strive to reproduce experimental structures rather than binding free energies. To do so, protein-ligand complexes are created according to simple atomic interaction-pair potentials. Among well-known approaches one can find the potential of mean force (PMF) [94–96] and DrugScore [97–99]. One of the weaknesses of these approaches is that they are based on a limited set of protein-ligand complexes, potentially making the applicability domain very narrow.

2.3.1.5 Rescoring

As mentioned above, mainly for speed efficiency reasons, the scoring functions of most docking tools are poorly evaluated, which ultimately yield a low hit rate [100]. In order to circumvent this problem, one can make use of *rescoring*. Usually *rescoring* is used as a secondary screening technique (refining stage), and one applies it to the top ranked molecules that were selected from a larger screening. This procedure evaluates more rigorously, for a given pose, the interaction between the ligand and the receptor. The estimation of the binding affinities are slower than standard scoring functions, but the accuracy will be significantly improved. This is afforded by more robust prediction models, such as the implementation of molecular mechanics–generalized Born surface area (MM-GBSA) techniques [100]. Many softwares were developed in order to refine the score, namely DrugScore [97–99], its improved version DSX [101] and HYDE [102].

2.3.1.6 Docking of fragments

Most of the docking programs were parametrized and optimized for drug-like molecules, making their utility dubious when it comes to evaluating smaller molecules such as fragments [22]. Consequently, several softwares were developed more specifically for fragment docking, including MCSS [103], SuperStar [104, 105] and SEED [77, 78].

Docking approaches combined with parallel computing allows one to virtually screen millions of compounds in few hours. This is a fast and cheap method to identify fragment that could act as starting points for ligand discovery efforts. However, fragments are small chemical entities they are likely to bind in different spots of the binding site. Therefore fragment docking can yield several poses per fragment. For instance, SEED exhaustively docks rigid fragments into the binding site and this procedure often yields more than a thousand poses per fragment. This number needs to be reduced to make the evaluation of the poses more manageable. A straightforward filter is the score associated with the pose. One usually only looks at the very best poses according to the score. However, Verdonk *et al.* [106] showed in a vast retrospective docking study that fragments were correctly docked only about 50% of the time. Interestingly, they also showed that this number could be improved by taking several poses into account. Many of the failed fragments were actually docked correctly but not scored high enough. This illustrates pretty well the current limiting factor in molecular docking: the scoring stage, which still poorly evaluates the free energy of a ligand binding to a protein. This can be explained by a lack of robust models to correctly estimate components of the binding free energy, such as desolvation terms for both ligand and receptor, water interactions, conformational entropy penalties and protein conformational changes [107]. Thus one often use docking as a tool to generate binding modes, rather than a binding affinities predictor.

As described in the following chapters, I developed different strategies to improve the results of fragment docking and make them more reliable. Chapter 4 illustrates how I used all the poses generated by SEED to create a map of potentially favorable interactions. Such maps were then filtered and applied to a fragment growing strategy in order to create ligands for the β_2 -adrenergic receptor. Chapter 6 and 7 illustrate how SCUBIDOO can be used to dock fragments disguised as drug-like compounds to make better use of most current docking software tools.

2.3.2 Ligand-based methods

“Small molecules with similar chemical properties should exhibit similar biological activities.” This is the famous pillar of medicinal chemistry which can simply summarize the

ligand-based approaches in the chemoinformatics field. These approaches require the knowledge of biologically active molecules in order to identify similar ones. A ligand-based screen usually compares two molecule libraries against each other. The chemical information contained in each molecule is numerically encoded using fingerprints, which consist of a finite bit string. The measure of similarity between two molecules is usually calculated using the Tanimoto coefficient, as defined in equation 2.5:

$$Similarity = \frac{N_{A \cap B}}{N_A + N_B - N_{A \cap B}} \quad (2.5)$$

where N_A and N_B are the numbers of bits in bit strings A and B, respectively, and $N_{A \cap B}$ is the number of bits which are common between the two bit strings A and B. The measure of similarity is between 0.0 and 1.0, where 1.0 indicates strict equivalence of the bit strings. Several fingerprints exist, each one encoding the chemical information differently. The MACCS fingerprint [29] and the extended-connectivity fingerprints (ECFP) [30] are broadly used fingerprints. The MACCS fingerprint contains 166 bits and is used for substructure search. Each bit position specifically encodes a common functional group. ECFP, in contrast, are topological circular fingerprints, which are not predefined and can represent a large number of different molecular features or substructures.

Ligand-based approaches are suitable for quickly retrieving analogs of known active ligands. Thus, these approaches are very helpful when no structural information about the target is known (Table 1). An illustration of the use of ligand-based approaches in an FBDD project is “*SAR by catalog*”, which will be described later in this chapter. Another recent application is the deconstruction-reconstruction approach [9], which is illustrated with SCUBIDOO and will be further detailed in chapters 5, 6 and 7.

2.3.3 Synergy of in silico methods

Structure-based approaches are more widely used in the drug discovery pipeline in comparison to ligand-based ones. However, these approaches are applicable in synergy to produce more robust outcomes [108, 109].

Ligand-based approaches can be utilized to pre-filter chemical libraries before structure-based approaches, such as docking, are applied. I illustrate this scenario in chapter 6, where we identified derivative products from SCUBIDOO based on known active scaffolds extracted from a database of known PIM-1 kinase inhibitor.

Another recent study by Martiny *et al.* [88], to which I contributed, highlighted how ligand- and structure-based approaches can be used in concert in order to predict the

inhibition of cytochrome P450 (CYP) 2D6. In this case, ligand-based methods are used to select, for a given targeted ligand, the most similar molecule among a set of known active ligands. Once the closest active is identified, the protein conformation associated with it is then selected as receptor for further docking of the original targeted ligand.

2.4 Experimental Methods

Several biophysical techniques allow for detection of fragment binding to a protein. Among them, surface plasmon resonance (SPR), isothermal calorimetry (ITC), thermal shift assay and mass spectrometry (MS) have been successfully implemented in fragment-based drug discovery [110, 111]. However, since the interactions remain weak, structural methods such as X-ray crystallography [112] and NMR [113] might be more accurate, leading to fewer false positives during screening campaigns [17]. These two structural techniques will determine the precise binding mode of the fragment within a protein target.

2.4.1 Surface Plasmon Resonance

SPR has become a standard biophysical technique for fragment screening due to the high sensitivity of the method as well as its cost effectiveness [114–118]. Indeed, only a small amount of protein is required (approximately 50 μg) for a screening campaign [119]. During an SPR experiment, the protein is tethered on the surface of a biosensor chip (typically, a glass slide covered with gold) and fragments are passed over it. A beam of polarized light is directed towards the metal surface and the change in refractive index is measured. These changes are correlated to the mass of the protein and the fragment. Binding events as well as kinetic information can be revealed using SPR. A number of successful SPR-based fragment screens have been reported on different targets, including the β_2 -adrenergic receptor [120], BACE-1 [121], MMP-12 [122], thrombin [123] and chymase [116].

2.4.2 Isothermal Titration Calorimetry

ITC is a biophysical technique that allows to measure precisely, at a given temperature, the difference in heat between two samples which are mixed together [124–126]. It is used to determine the thermodynamic profile of a ligand upon binding to a protein. ITC measures the free energy of binding ΔG (affinity) and the enthalpy ΔH , which can then be used to calculate the entropy ΔS and the reaction stoichiometry (n). Having access

to such profiles makes ITC certainly a powerful approach for ligand optimization, since one can determine the enthalpic and entropic components of the binding energy. This is extremely valuable in order to focus on fragments that are more enthalpic binders. Those fragments are believed to be better starting points for optimization than entropic binders [127]. However, due to the fact that assays require higher protein concentration, ITC is not yet suited for large scale screening, making it a second or even third screening technique rather than a primary one [128]. A good illustration of the use of ITC in fragment optimization is the work of Edink *et al.* [129]. In this study, the authors successfully grew a fragment towards an efficient ligand of acetylcholine-binding protein (AChBP).

2.4.3 Thermal Shift Assay

TSA monitors the change in thermal stability of a protein under various conditions (e.g. ligand concentration, pH or salts) [130–134]. At a certain temperature, the protein will unfold and its hydrophobic core will be exposed and gets into contact with a sensitive fluorescent dye. Ligands that bind to the protein would be expected to increase protein thermal stability (i.e. the unfolding temperature) [135]. TSA is considered one of the quickest and easiest methods for fragment screening [25, 107]. Due to the weak interactions of fragments, the thermal shifts are expected to be small. A fragment is considered to be a hit when the shift in temperature is above 1°C (i.e. twice the standard deviation of the measured melting point). TSA was successfully applied to the discovery of fragments binding to the Y220C p53 tumor suppressor protein [136]. This technique was also successfully applied in chapter 6 for the identification of a potentially fragment-sized PIM1 binder.

2.4.4 High Concentration Screening

Biochemical assays are not the best-suited for fragment screening due to their poor ability to detect fragments that bind weakly. Screening at high concentrations usually yields a higher number of false positives due to reactivity, aggregation or interference with the assays [118, 137–139]. Additionally, high concentrations require high solubility of the fragments (about 1 mM). Although HCS has limitations, it offers the advantages of HTS (i.e. mainly high throughput nature, low protein consumption and wide applicability) [117]. Despite the aforementioned limitations, HCS was successfully applied as primary fragment screening technique to identify inhibitors of beta-secretase (BACE-1) [140], HSP90 [141], Checkpoint Kinase 2 (CHK2) [142] and B-raf [143]. The latter

yielded Vemurafenib three years later, the first FDA-approved drug resulting from an FBDD effort.

2.4.5 Mass Spectrometry

Electrospray ionization mass spectrometry (ESI-MS) is an analytical method which consists of ionizing a molecule and measuring its mass-to-charge ratio in gas phase in order to detect its molecular weight. This technique can be applied to detect the binding of fragments to a protein (i.e. an increase of the mass of the protein), and allows one to calculate the dissociation constant (K_d) as well as the stoichiometry [144]. ESI-MS requires only a small amount of protein or ligand and is considered as a fast primary screening technique [145, 146]. The use of ESI-MS as fragment screening technique is not as frequent as other methods, but nonetheless success stories exist for Hsp90 [144, 147] or endothiapepsin [25].

2.4.6 Nuclear Magnetic Resonance

FBDD owes a lot to NMR due to the “SAR by NMR” in 1996 which was the first success of a FBDD effort [148]. NMR is a powerful technique for 3D structure determination as well as the measurement of protein-ligand interactions [149, 150]. Briefly, NMR spectroscopy rests on the magnetic properties of several atomic nuclei (N,C,H,F) to give further information about the chemical environment they belong to. Two main approaches exist for NMR screening: protein-detected NMR and ligand-detected NMR [151–153]. The protein-detected NMR can detect mM to nM interactions and provides information about the binding site. However, it requires a large amount of labeled protein (50-200 mg) with high solubility [154]. Ligand-detected NMR can be done using different techniques, most notably saturation transfer difference (STD) [155, 156] and water-ligand observed by gradient spectroscopy (WaterLOGSY) [157, 158]. In those methods, an irradiation pulse is applied at the resonance frequency of the biomolecule (STD) or the water bulk (WaterLOGSY). The relaxation properties of the fragments extracted from the ^1H NMR signals allow to distinguish bound from unbound fragments. NMR is often used as experimental technique in FBDD and was applied successfully to many targets, mainly Hsp90 [159], Bcl2 [160], BACE-1 [161], PDK1 [162].

2.4.7 X-ray Crystallography

X-ray crystallography is a diffraction method which can provide very high resolution data of a macromolecule (crystal) down to atomic level [163–166]. Crystallography is

essential in the process of ligand optimization and is the pillar of any structure-based campaign [167, 168]. This technique consists of bombarding the crystal lattice atoms by X-ray beams. The diffraction information will then be used to generate a three-dimensional picture of the electron densities. In the context of protein-ligand crystallography, two main approaches are commonly used: soaking and co-crystallization. Soaking consists of preparing pools of diverse fragments (each pool usually contains up to 10 fragments) and soaking the pools into the apo-protein crystals. Soaking can be done with single fragment, but fragment pooling is believed to speed up the process. This approach allows the structural biologist to identify (multiples) bound fragments in the protein pockets [169]. Within the crystal, the protein ought to have a solvent-exposed or unhindered binding site. If this is not the case, one can turn to co-crystallization in order to determine the protein-ligand complex. The latter approach consists of crystallizing a protein and a fragment. X-ray crystallography offers many advantages, most notably the protein-ligand structure which can help to a better understanding of the binding mode. It also gives the possibility to improve a crystallized fragment to a more potent one using linking, merging or growing approaches (cf fragment optimization). X-ray crystallography yields ligands with validated binding mode which make this technique reliable [117] as secondary screening technique.

Nevertheless, a recent study done by Schiebel et al. from the Klebe group suggested crystallography to be used as a primary screening technique in fragment-based lead discovery efforts [26]. In a previous paper [25], the authors screened a library of 361 fragments [27] against endothiapepsin using six different biophysical screening methods, namely fluorescence-based HCS, STD-NMR, RDA, native MS, MST and TSA. The outcomes showed that the hits mutual overlap was surprisingly low. The authors decided to have a closer look, and exhaustively soaked all the 361 fragments individually into the protein. This procedure yielded 71 structures (20% hit rate !). Alarmingly, 31 hits (44%) were not detected by the six aforementioned screening techniques and only 21 hits (30%) were detected by one method. Hence, the best combination of two screening techniques would only have yield a 26% hit rate. While this study surely emphasize the effectiveness of crystallography, it is important to mention the described effort was an Herculean task. Indeed, the attempt to crystallize 361 fragments took several months involving many experts and a frequent access to a synchrotron beam. All those conditions are not yet within the reach of everyone, but likely with enhancement of synchrotron source this situation will change.

However, X-ray crystallography is still a low-throughput screening technique, it does not provide the affinity of crystallized ligands and it is limited to protein targets that are susceptible to crystallization (GPCRs are not amenable to such screening, as discussed in chapter 3). X-ray crystallography was implemented in numerous fragment-based

efforts aiming at different targets: endothiasepsin [25, 27], cyclin dependent kinase 2 (CDK2) [170], aurora A kinase [171], Janus kinase 2 (JAK-2) [172], PIM-1 [173] and thrombin [174].

2.5 Synergy between in silico and in vitro Methods

Experimental techniques allow to identify fragment hits. Computational approaches such as docking will help to predict fragment binders, but will ultimately need confirmation through experimental assays. In the best course of action, X-ray crystallography will be used as ultimate step in order to determine the precise binding mode of the fragment. This will be crucial for a later optimization (i.e. make the fragment bigger and more potent). Should X-ray determination of the structure fail or be inapplicable, one can still rely on computational approaches (docking) to predict the binding mode of the fragment hit. This option is illustrated in chapter 4, where we successfully grew new potent ligands for the β_2 AR based on the predicted binding mode of fragments. Computational approaches can also assist X-ray experiments when preparing the pool of fragments or help to prepare a fragment library for screening. In both cases, chemoinformatic approaches are implemented in order to ensure chemical diversity among the selected fragments, as well as an estimation of the solubility if it is not available experimentally. Diversity can be defined using standard Lipinski descriptors, any fingerprint or more complex 3D descriptors such as shape complementarity (ROCS) [175, 176].

2.6 Fragment Optimization

Once the binding mode of a fragment has been described, either experimentally or computationally, it can be used as starting point for an FBDD effort. Optimization can then follow, striving at expanding the fragment into a bigger molecule and increasing its potency and specificity. Three main approaches are conceivable in order to extend fragments: growing, merging and linking (figure 2.5).

2.6.1 Growing

Fragment growing is the most often applied approach because it is considered more straightforward and successful than the other techniques [177–179]. Growing aims at extending a promising fragment (i.e. *core* fragment, usually the one that makes the most compelling interactions) by adding new chemical features that will make additional interactions within the binding site. Ultimately, doing so will improve the initial potency.

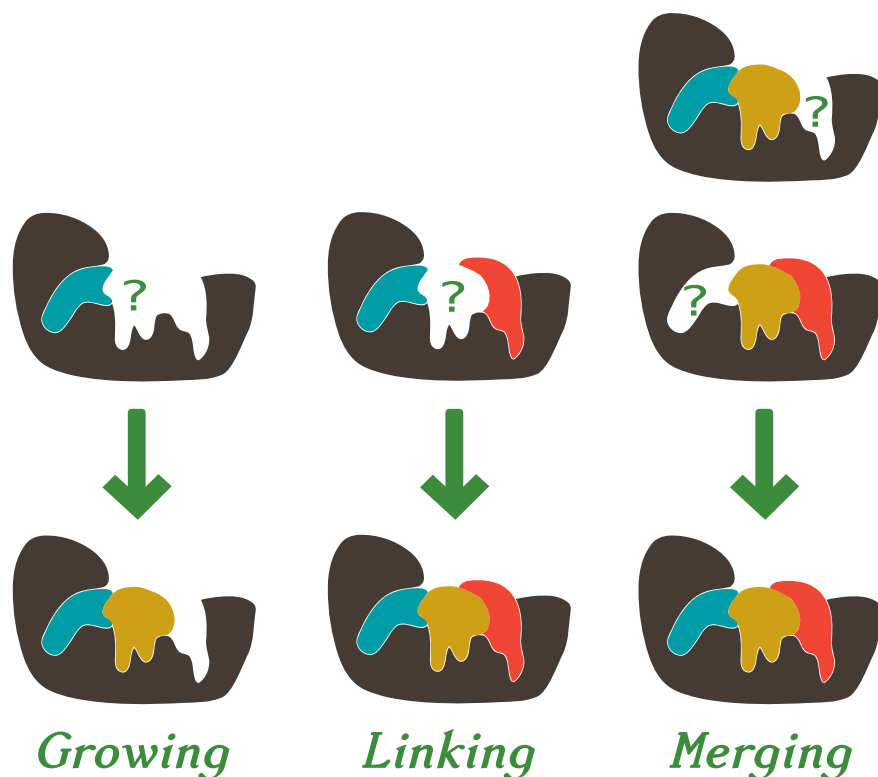


FIGURE 2.5: Fragment based approaches in drug discovery: growing, linking and merging.

A compatible candidate for growing should contain reactive chemical features in an unobstructed site. Depending on the chemical feature, one can then choose an organic reaction in order to attach another group to the fragment. This scenario is illustrated in chapter 4, where we successfully grew new ligands for the β_2 -adrenergic receptor using reductive amination. One of the key considerations during the growing procedure is that the binding mode of the *core* fragment should remain the same, even after several growing iterations.

Fragment growing strategies have been applied successfully to various targets [118, 180, 181] including BACE1 [182], acetylcholine-binding protein (AChBP) [129], matrix metalloproteinases (MMP) [183] and phosphatidylinositol-3 kinases (PI3Ks) [184].

Growing approaches are extensively supported by computers and a lot of software tools are available, including LUDI [89], SPROUT [185], CONCERTS [186], ReCore [187], Caveat [188], BREED [189], GANDI [69] and BROOD [190]. However, one of the main challenges of these programs is to assure synthetic feasibility. Several rules can be implemented in order to assess whether or not a compound is synthetically feasible, among them the most often used being BRICS [191], RECAP [192] or an estimation of the so-called synthetic accessibility [193]. These approaches are based on knowledge of

chemical reactions in order to fragment the molecules. However, this does not assure that the resulting fragments exist. An example of this problem is GANDI. This software allows to link or grow fragments that were previously docked with SEED. A library of linkers can be defined, but ultimately the process of joining two fragments together is based on heavy atom (i.e. non hydrogen) bonds. For instance, if two carbons of different fragments are in reasonable distance and angles to each other, the two atoms are linked. Then, a simple minimization usually allows one to see if the generated compound is energetically favorable. Should this be the case, the problem is still that the created compound rely on the formation of a C-C bond, which is not the simplest connection to achieve in organic chemistry. Therefore, even though most of the generated compounds look appealing *in silico*, they might still be far from being synthesizable.

In order to bring a solution to this problem, we developed PINGUI, an *in silico* structure-based growing workflow based on robust organic reactions and available building blocks. The creation and application of PINGUI is illustrated in chapter 4.

2.6.2 Merging

Merging fuses two fragments that contain a common chemical feature known to bind in the same position in a given target. The procedure is far from being easy to achieve, mainly due to conformational limitations which need to be overcome. However, should the merging be successful, the loss of translational and rotational degrees of freedom of the two fragments is usually highly entropically favorable (up to 20 kcal/mol). A recent study from Hudson *et al.* [194] illustrates this challenge and also offers guidance on how merging could be done. The authors observed that the success (or failure) of merging is correlated to the average distance between the common atoms of the two initial fragments (i.e. not fused yet). Their analysis estimates that the maximum average distance should be below 1Å. Moreover, the authors highlight the importance to correctly estimate the internal strain of the generated molecule. If the two fused fragments are able to reproduce their initial binding mode, but only with strong conformational constraints, the generated molecule is likely to fail.

Merging strategies remain rarely applied in FBDD efforts [107], but some success stories exist for targets including Hsp90 [159], Jun N-terminal kinase 3 (JNK3) [195] and PI3γ kinase [184].

2.6.3 Linking

Linking describes the process of joining together two non-competitive fragments (i.e. fragments that bind in two different positions of the binding site) with a linker. Theoretically, the generated compound should offer a lower free energy of binding than the sum of the two initial fragments [107, 196]. The process of linking might be considered as the most challenging fragment optimization approach. Indeed, it requires to find a linker which will respect the original conformational constraints of the two initial fragments while making favorable interactions with the protein. On top of that, the linker should introduce only little entropic penalty (i.e. avoid too many rotors) and synthetic feasibility restrictions might arise, limiting the number of compatible linkers even further.

Albeit being challenging, linking strategies were flourishingly applied to different targets including thrombin [174], FK506 [11], lactate dehydrogenase [197] or Hsp90 [198].

One way to tackle the aforementioned challenges would be to divide the linking process into two approaches we are already familiar with: growing and merging. Indeed, linking could be assimilated as a two-step growing (A towards B and C towards B) coupled with a merging (A-B with B-C) as illustrated in figure 2.6. Therefore, SCUBIDOO and PINGUI, which were applied successfully in a growing context, could also be used as a tool for a linking project.

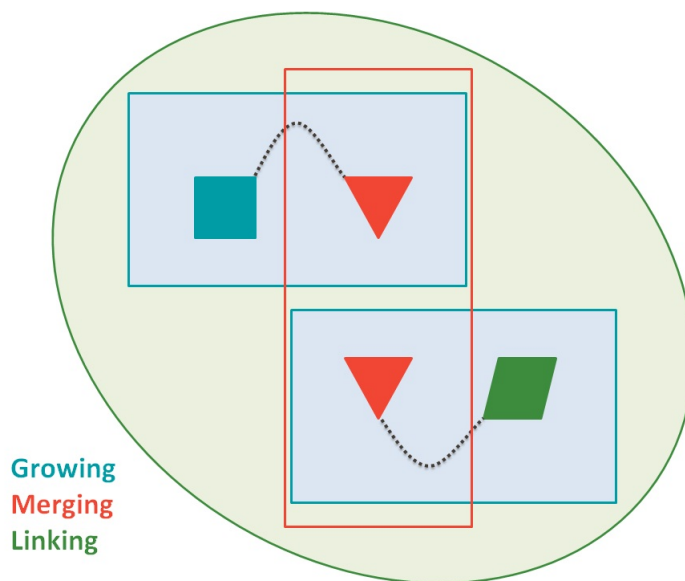


FIGURE 2.6: Linking could be assimilated as a 2-step growing A (blue square) towards B (red triangle) and C (green rectangle) towards B coupled with a merging (A-B with B-C)

2.6.4 SAR by Catalog

SAR by catalog consists of searching for similar compounds or compounds that do contain the fragment of interest in databases of commercially available molecules [199]. This is usually done by similarity search using common fingerprints. Once analogs have been identified, they can be purchased and tested. This approach can be used to quickly generate derivatives for a fragment of interest. However, the number of generated molecules is usually quite limited and this approach relies on already existing molecules. A good illustration of this approach is the work of Jahnke *et al.* [200] at Novartis where they designed new allosteric non-bisphosphonate FPPS inhibitors .

2.7 Conclusions

With currently two approved drug and more than 30 clinical candidates [201], fragment-based approaches are gaining strong momentum in drug discovery efforts. Success stories have been flourishing for a wide range of targets including challenging ones such as membrane receptors (GPCRs). Those campaigns yielded high quality lead compounds based on the identification of fragments paired with high LE.

This accomplishment rests on the recent progress made in the experimental field, where fragment screening approaches made considerable leaps forward. Every experimental method has its pros and cons that must be carefully weighted depending on the target. Several techniques can be used in parallel to improve the reliability of the prediction. X-ray crystallography and NMR are still the most reliable techniques which can be used in the final step, since they allow to depict the fragment binding mode in its environment. This is crucial for further optimization. Despite many advantages, fragment screening approaches lack high-throughput format and often require high amounts of fragment and protein.

This low-throughput nature and compound requirement can be compensated by computational approaches, which allow to virtually screen millions of fragments in a few hours. Thus, fragment libraries can be pruned in order to select only fragments that are predicted as binders. This synergy allows to pre-select promising fragments and potentially increase the hit rate in experimental assays. *In silico* approaches are also powerful at generating plausible extensions when it comes to the fragment optimization stage (growing or linking). Several tools were developed in order to assist the optimization, but they often lack reliable synthetic feasibility filters, thus suggesting molecules that might be impossible to synthesize with reasonable effort. Moreover, fragment docking remains challenging because most of the presently available tools were developed for

drug-like (bigger) compounds. Although several tools were developed to address this issue, fragment binding mode prediction remains challenging due to the small nature of fragments that tend to bind in different spots.

To address these two critical problems, I developed two tools in order to contribute to the fragment elaboration cycle (figure 2.1). PINGUI (chapter 4) is an *in silico* workflow aiming at growing fragments in a structure-based fashion combined with an assessment of the synthetic feasibility. SCUBIDOO (chapter 6) is a free online database currently containing 21M compounds (most of which do not exist yet) optimized towards high likelihood of synthetic tractability. Every product in this database is the assembly of two commercially available building blocks and comes with synthetic instructions. Docking SCUBIDOO products could be equivalent to docking fragments disguised as drug-like compounds. This notion will be explained in more detail in chapter 6, 7 and 8.

"The man who passes the sentence should swing the sword."

Lord Eddard Stark, *A Game of Thrones* (1996)

The Chapter 3 is made of two parts. A first part, written by myself, gives an introduction to the GPCRs. The second part is an article which was submitted to the *Journal of Medicinal Chemistry* and it is currently in revision. The authors list is the following (by contribution order): Sylwia Gawron, Tonia Aristotelous, Florent Chevillard, Paul G. Wyatt, Andrew L. Hopkins, Peter Kolb, Iva Hopkins Navratilova, Ian H. Gilbert. Sylwia handled the synthesis part and the overall strategy. Tonia and Paul were responsible for the SPR measurements. I contributed to the docking part, generation of the pictures and binding modes analyses.

Chapter 3

Structure-free optimisation of fragments for the beta-2 adrenergic receptor

3.1 Introduction to GPCRs

G protein-coupled receptors (GPCRs) are the largest class of membrane proteins in the human genome. They are involved in many physiological pathways including visual or olfactory senses, mood regulation, inflammation and immune system regulation. Because of this important role they are targeted by approximately 40% of modern drugs [202–204]. GPCRs are also known as seven-transmembrane domain receptors (7TM receptors), because their structure is composed of seven helices that pass through the plasma membrane.

GPCRs transmit signals from the cellular environment to the cytoplasm upon ligand binding. The ligand binding process takes place in the extracellular environment and will induce conformational changes in the receptor structure that will be translated into a biological function. Some GPCRs exhibit a basal signaling activity that is modified upon ligand binding. Agonists increase the basal activity while inverse agonists decrease it. Ligands that do not alter the basal activity are called antagonists and block the access to the orthosteric site (i.e. the cavity where the endogenous ligand binds), thus preventing receptor activation. In an inactive conformation, the intracellular part binds to a nearby G protein, which consists of three subunits G_α , G_β and G_γ . When GPCRs are activated, the intracellular part of the receptor will undergo major changes, prompting the G_α subunit to exchange guanosine diphosphate (GDP) with guanosine triphosphate (GTP). This exchange will provoke the dissociation of G_α from the remaining $G_{\beta\gamma}$ complex and

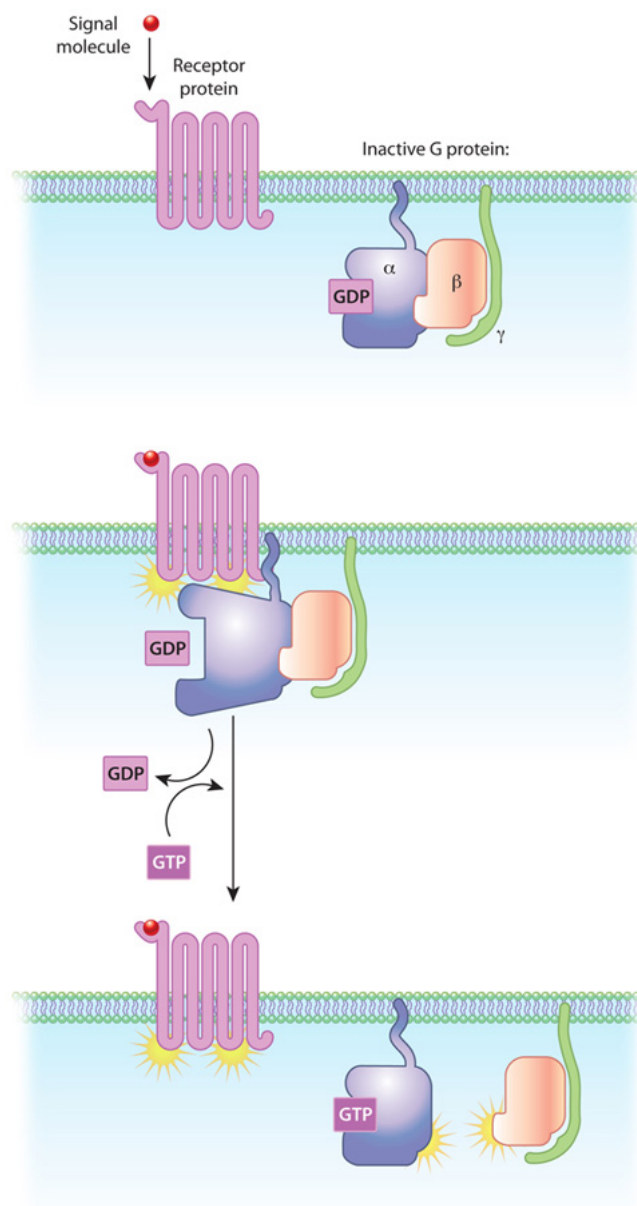


FIGURE 3.1: GPCR activation. Reprinted from Nature Publishing Group. Li, J. et al. The Molecule Pages database. Nature 420, 716-717 (2002). All rights reserved.

the receptor (figure 3.1). G_{α} will trigger downstream cascades, resulting in different biological signals involving other cytoplasmic proteins.

The β -adrenergic receptors (β AR) are an important subfamily of the GPCR and by extension, relevant drug targets. β AR have three subtypes β_1 , β_2 and β_3 . Activation of the β_2 AR triggers bronchodilation, and agonists are crucial drugs for the treatment of asthma. On the other hand, blocking the activation of the β_2 AR results in nullifying the effect of the endogenous ligand epinephrine, which is involved in the *fight-or-flight* response or more commonly known as stress response. Antagonists are medically used to prevent tachycardia, cardiac arrhythmia or hypertension.

GPCR ligands are identified using experimental assays, which can be divided in different categories, namely receptor binding assays, G protein-dependent functional assays, G protein-independent functional assays and receptor dimerization assays. Each category will be briefly discussed below.

Receptor binding assays are used to identify GPCR ligands in a receptor-containing membrane environment [205]. These assays can be used to determine the interactions between a receptor and its ligand, namely the dissociation constant or the affinity. Such assays are cell-free methods, which means they do not involve any downstream signaling after the receptor. Consequently, the limit of these methods is that they do not provide information about the efficacy profile of the ligand (agonist, inverse agonist or antagonist). Radioligand binding assays consist of displacing the binding of a radiolabeled (3H or ^{125}I) ligand to the receptor by a non-labeled compound. Such methods are high throughput, but they do rely on the availability of radioligand (often expensive) and generate radioactive waste.

G protein dependent functional assays analyze the biological response within a cell after ligand binding. As mentioned earlier, upon ligand binding, the receptor will undergo major conformational changes, which will activate the coupled G proteins. Once activated, G proteins will triggers additional downstream signals. Several assays were developed to measure G protein activation or G protein-mediated signals, namely GTP γ S binding assays, cAMP assays and Ca^{2+} or $IP_{1/3}$ accumulation assays (figure 3.2).

- GTP γ S binding assays measure the amount of non-hydrolyzable GTP analogue to the $G\alpha$ subunit [206].
- cAMP assays measure the intracellular concentration of cAMP and rely on the activity of adenylyl cyclase, which is regulated by $G\alpha_s$ or $G\alpha_{i/o}$ proteins [207]. The activation of $G\alpha_s$ stimulates adenylyl cyclase activity, yielding an increased level of cellular cAMP, while the activation of $G\alpha_{i/o}$ yields a decrease of cAMP level.
- Ca^{2+} or $IP_{1/3}$ accumulation assays rely on the activation of $G\alpha_q$ or $G\alpha_i$ which activates phospholipase C (PLC), which in turn hydrolyzes phosphatidylinositol biphosphate (PIP_2) to form inositol triphosphate (IP_3). IP_3 activates the IP_3 receptor, resulting in an outflow of Ca^{2+} from the endoplasmic reticulum (ER) to the cytoplasm and thus an increasing intracellular Ca^{2+} concentration [208]. IP_3 is quickly hydrolyzed to IP_2 and then to IP_1 .

GPCRs may signal independently of G proteins. In order to monitor this, one needs to make use of internalization assays which rely on the desensitization of GPCRs [209].

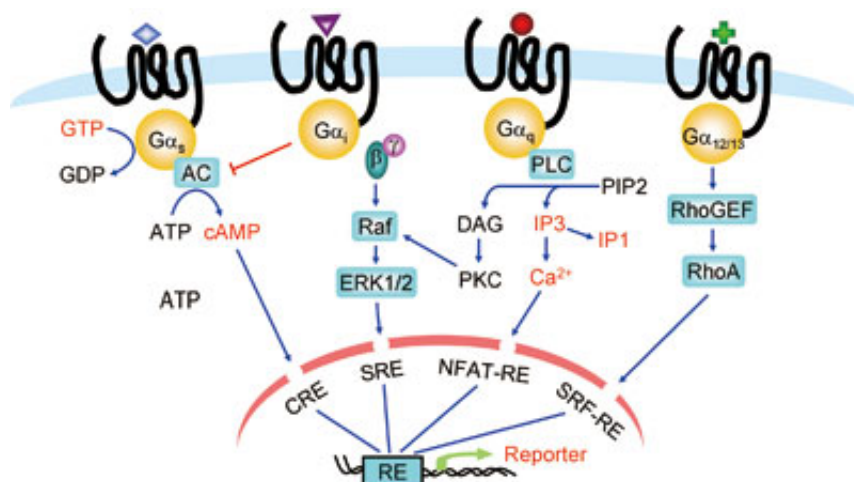


FIGURE 3.2: G-protein-dependent assays. Reprinted from Nature Publishing Group. Zhang, R. et al. Tools for GPCR drug discovery. *Acta Pharmacologica Sinica* (2012) 33: 372–384. All rights reserved.

During this process, GRKs phosphorylate activated GPCRs. The newly phosphorylated GPCR will bind to cytosolic β -arrestins, thus preventing the activation of the G protein. G protein independent functional assays consist of different methods, one of the most often used is the β -arrestin recruitment assay [210]. The β -arrestin pathway is distinct from G protein pathway, meaning that each pathway could be modulated via 'biased ligands' [211]. Such ligands are pharmaceutically crucial, since they could potentially give more insights into the functional selectivity of GPCRs. Indeed, this could help to understand and potentially avoid side effects by activating (or suppressing) specific pathways [212].

GPCRs can form dimers or oligomers with other GPCRs [213] and the dimerization process can be monitored using various techniques. Among those techniques, the Foerster resonance energy transfer (FRET) approach is widely applied [214–216]. FRET techniques rely on the energy transfer between two chromophores (an acceptor and a donor). The donor will transfer energy when an acceptor is in close vicinity. In a GPCR context, donor and acceptor chromophores are fused to the C-terminus of GPCRs, and energy transfer will be observed when donor and acceptor are brought close to each other (i.e. dimerization). Dimerization of GPCRs has a crucial impact on receptor pharmacology and signaling [217]. Thus, ligands targeting dimers or blocking dimer formation may have very specific therapeutic effects [218].

Despite being important drug targets, crystal structures of GPCRs have been rare. This can be explained by the difficulty of stabilizing the receptor in a membrane-like environment. Furthermore, GPCRs are oscillating between inactive and active state conformations and contain highly flexible intracellular and extracellular loops (ICL and ECL respectively), hampering the crystallization process even more [219].

In 2007, Cherezov *et al.* were the first to successfully crystallize human GPCRs, with the β_2 AR in an inactive conformation [220]. To do so, they engineered a β_2 AR structure by replacing the highly flexible ICL3 with T4 lysozyme [221], allowing to reduce conformational heterogeneity. This breakthrough opened many doors in both experimental and computational fields. For instance, many opportunities flourished in structure-based approaches such as docking, with the possibility to virtually screen the β_2 AR with a better accuracy than the current receptor models. This was quickly accomplished by Kolb *et al.* [222] in 2009, with the discovery of novel chemotypes binding to the β_2 AR as well as a high-affinity inverse agonist ($K_i = 9$ nM). A year later, the high affinity ligand was crystallized [223] and the structure revealed that the predicted binding mode was very close to the resolved one (RMSD = 0.9 Å).

However, inactive structures do not give information about the activation of the receptor, which is also critical for a better understanding of GPCRs. The gap was filled in 2011 by Rasmussen *et al.* [224] with the resolution of an active-like conformation of the β_2 AR in complex with an agonist (BI-167107). In this case, the challenge was to mimic the activated G protein binding to the intracellular part. This was achieved by engineering a nanobody (Nb80) acting as a surrogate of $G\alpha$. In a companion manuscript, Rosenbaum *et al.* showed that even when bound to a covalent agonist, the β_2 AR crystallizes in an inactive conformation [225]. This indicated that G protein or nanobody interaction is required to stabilize the active conformation. These studies were followed by the determination of active conformations in complex with three different agonists, including epinephrine [226], an improved nanobody (Nb6B9), as well as a covalent ligand complex [227].

With the knowledge of snapshots of both active and inactive conformations of the β_2 AR, insights into the activation mechanism started to be revealed. The biggest difference was observed in the cytoplasmic part, where the extremity of helix 6 shifted 11 Å outwards in order to accommodate the activated G protein. Interestingly, no such drastic changes were observed in the binding site. The major change is manifested mostly in His296^{6.58} (figure). In the active receptor conformation, His296^{6.58} is part of a large polar network involving the catechol moiety of the ligand, a water molecule, Tyr308^{7.35}, Ser204^{5.43} and Asn293^{6.55}. This polar network is located close to the orthosteric site, making it smaller in the active conformation. In the inactive conformation of the receptor, the aforementioned polar network is not so extended, resulting in an outward shift of His296^{6.58} [226, 228], resulting in a larger orthosteric site (Figure 3.3).

In parallel, other members of the GPCR family were successfully crystallized, including the turkey β_1 AR [229], adenosine A_{2A} [230], dopamine D₃ [231], CXCR4 [232], histamine H₁ [233], M_2 muscarinic receptor [234] or more recently OX₂ orexin receptor [235].

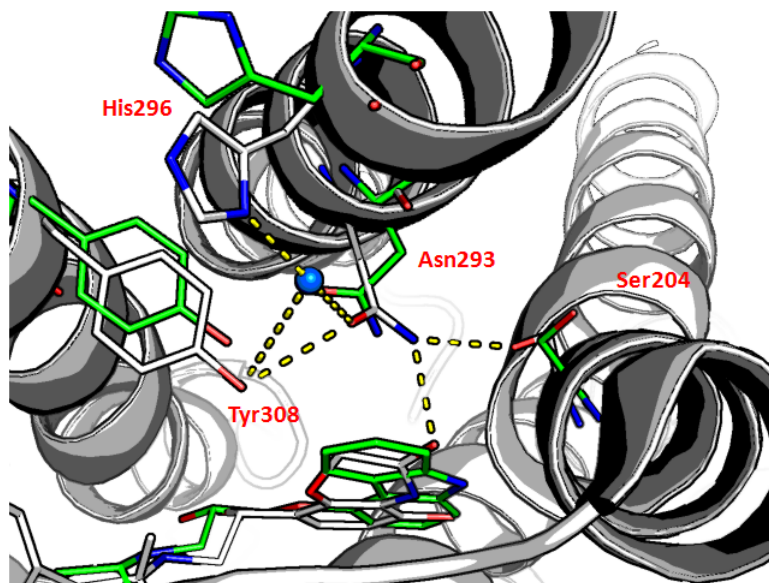


FIGURE 3.3: Polar network involved in the activation of the β_2 AR. The receptor in its active is depicted in white carbons, while the corresponding residues of the receptor in its inactive conformation are depicted in green carbons.

Despite the fact that a sizeable number of GPCR structures have been crystallized in both active and inactive conformations, the mechanism of activation remains poorly understood [226]. Every new structure solved can be seen as another snapshot of a stable conformation which brings a new piece to the big puzzle that is the GPCR family. In the future, diversity of the yet-to-be crystallized ligands should be emphasized in order to address a maximum of key residues that could be involved in GPCR activation. Such tasks can be assisted by *in silico* techniques and will be illustrated in chapter 7, where we started to generate tailored molecules that could bind to a multitude of polar residues in proximity of the β_2 AR orthosteric binding site.

3.2 Abstract

Fragment-based drug discovery is making a major impact in the development of new therapeutic agents. Generally this requires co-crystal structures of the protein and fragment to guide the ligand optimisation. This is as there are multiple possible vectors along which the fragment can be optimised and it is difficult to predict the optimal vectors in the absence of structure. In this paper, we describe a protocol to optimise fragments in the absence of structure. We have developed a decision tree to guide optimisation of both the core scaffold and the different possible vectors for substituents. This was applied to optimisation of fragment hits of the β_2 -adrenoreceptor (β_2 AR), allowing generation of a “virtual” model of the active site. The decision tree provided a

framework for optimisation of the fragments. Subsequently, it was possible to rationalise the data by docking the fragments into a crystal structure of the receptor.

3.3 Introduction

Over the past 15 years, Fragment-based Drug Discovery (FBDD) has become an important and commonly used strategy for discovering small molecule drugs as evidenced by delivery of drugs into the clinic and others in clinical development [236]. There are several key benefits of FBDD. Firstly, it is possible to cover chemical space with far fewer molecules than when using a conventional “lead-like” or “drug-like” chemical library, as the molecules are much less complex [107]. Analysis by Fink and Reymond [237], suggests that each additional heavy atom in a molecule increases the size of the chemical space accessible by known synthetic chemistry by 8-fold. It can thus be concluded that screening a 1000 member library that averages 14 heavy atoms (190Da) would be equivalent to screening a library of over 1018 molecules of 32 atoms (450Da). Secondly, as the fragments are small, they can often form very ligand-efficient interactions [14] with the protein target, and it is more likely that they can take up optimum binding interactions with the binding site [16]. When optimising the compound, it is very important to maintain these optimum binding orientations of the ligand to the protein to retain the ligand efficiency (LE). In the vast majority of cases, this optimisation uses a Structure Based Drug Discovery (SBDD) approach, relying on repetitive co-crystallisation of ligands as they are optimised [179]. Unfortunately this limits the scope of FBDD, as many drug targets cannot be crystallised, or are not amenable to “soaking” or rapid repetitive co-crystallisation with ligands. Without knowledge of the binding mode of a novel hit, it is very difficult to prioritise amongst a very large number of possible analogue design ideas, given multiple vectors which can be optimised. In some cases computational approaches will be useful in guiding optimisation in the absence of 3-dimensional structures, but in general modelling methods are much more powerful and predictive when they are guided by experimental binding modes. In this paper, we describe an approach to fragment optimisation without the structural information available. To facilitate this, we developed a decision-tree (or work flow) to try and minimise the number of compounds that we needed to make and to maximise the information from each compound (Figure 3.4). There are two parts to this process: core optimisation and vector optimisation. During the initial core and vector optimisation, maintaining good ligand efficiency was a key goal. The fragment hits undergo core optimisation to maximise interaction of the scaffold. The core can be optimised by scaffold hopping, probing the effect of saturation/ unsaturation on the activity, isosteric replacement of functional groups, and changing the atoms in heterocycles. Initial optimisation will be driven by a combination

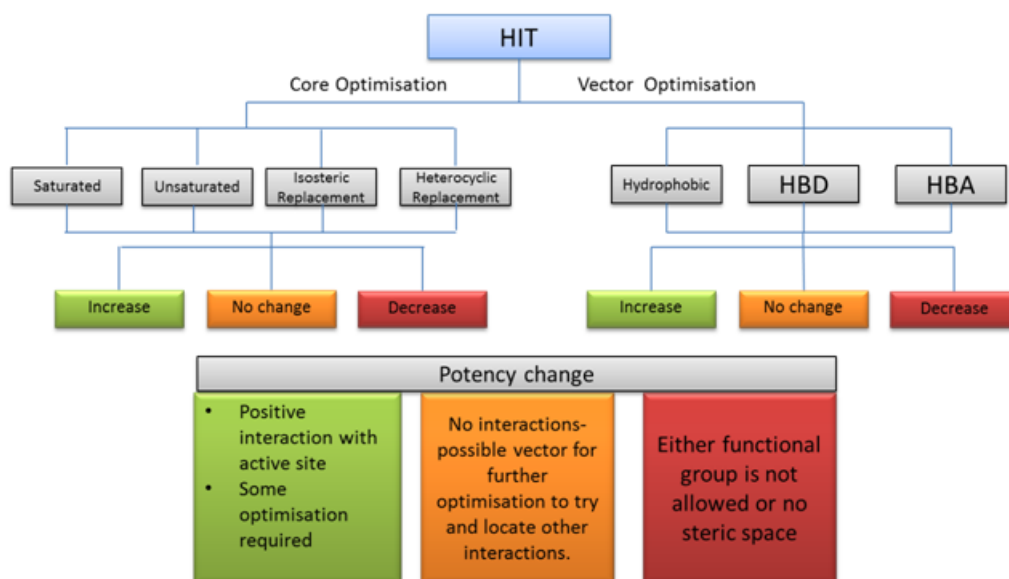


FIGURE 3.4: Decision tree

of ligand efficiency as well as overall potency, whilst retaining drug-like physicochemical properties. Vector optimisation will be achieved by probing each possible vector with a hydrophobic probe (e.g. a methyl) and a hydrogen-bond donor/acceptor which are also polar (e.g. a hydroxyl group). Where the fragment hit is binding tightly to the protein binding pocket, any probe at this position (either methyl or hydroxyl) will reduce potency. Where the probe has no effect on activity at a position on the fragment, it implies that such a position is not immediately adjacent to the protein binding pocket, that this vector points into solvent or that it does not interact directly with the protein. This case suggests further probing of this position would be valuable to see if additional interactions can be picked up. Finally, if the probe causes an increase in binding potency, this suggests a favourable interaction with the binding pocket. Further work is then required to see if this is an optimum interaction or whether there is more scope for optimisation.

3.4 GPCRs- beta receptor

G-protein coupled receptors (GPCRs) represent the largest family of membrane-bound receptors. They are key modulators of cellular processes including inflammation, secretion, neurotransmission, cellular metabolism, and growth. In consequence, GPCRs are one of the most common drug targets among the current pharmacopoeia [238]. The β_1 , β_2 and β_3 adrenergic receptors are class A GPCRs, activated by endogenous ligands adrenaline and noradrenaline [239]. Drug discovery efforts have produced a number of clinically relevant agonist and antagonist molecules. For example, agonists of the

β_2 AR present in cardiac, respiratory tract smooth muscle and adipose tissue are used to treat asthma. Historically, GPCR ligands have been discovered by high-throughput screening of cellular systems [240]. Recently, Surface Plasmon Resonance (SPR) an important technology for fragment screening [119], has been extended to GPCRs [241]. The advantage of SPR based screening is that it has a potential to identify orthosteric and allosteric ligands of proteins and can also use protein complexes as targets. This opens up new possibilities, as crystallography of GPCRs is not yet amenable to the medium- or high-throughput approaches necessary for ligand design [223?]. Recently we demonstrated a FBDD approach to discovering GPCR ligands by SPR screening, which resulted in highly potent fragments towards the β_2 AR wild type receptor [120]. In this paper, we report a structure-free optimisation of β_2 AR receptor ligands using SPR screening of wild type receptor as the experimental readout.

3.5 Chemistry and Biology

Our starting point as indicated in our original publication was a set of 4-substituted quinoline analogues (Figure 3.5) [120]. A total of five fragment hits were reported: fragments A to E with dissociation constants ranging from $K_D = 0.017 \mu\text{M}$ to $K_D = 22 \mu\text{M}$. These were active through binding to the orthosteric pocket. Functional studies indicated that these acted as antagonists through binding to the orthosteric site. A similar chemotype has also been reported by Christopher *et al* [242]. Compounds were synthesised using 4-chloro quinoline derivatives which reacted with amines in presence of K_2CO_3 as a base in DMF (Figure 3.6). Compounds were assayed using SPR; this was a powerful and rapid way to screen a focused set of compounds, giving information including binding affinity, and “on-” and “off-rates” (Table 3.15). It is important to note the SPR assays were conducted at 10°C to increase stability of the receptor on the sensor surface. Affinities may appear to be higher and off-rates slower at lower temperatures, however the data collected using SPR were used for compound ranking. Compounds generally had a moderate residence time and the compound with highest affinity (24) had the longest residence time.

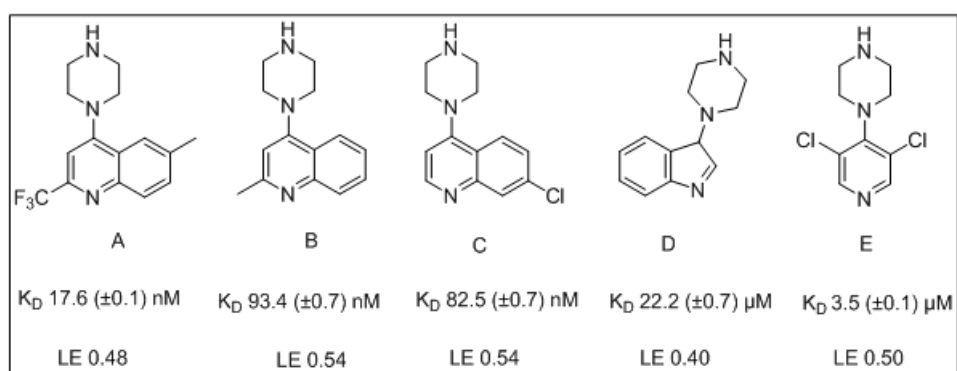


FIGURE 3.5: The initial hits against the β_2 AR

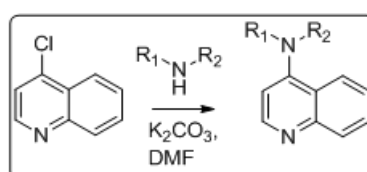


FIGURE 3.6: Synthesis of the quinolines

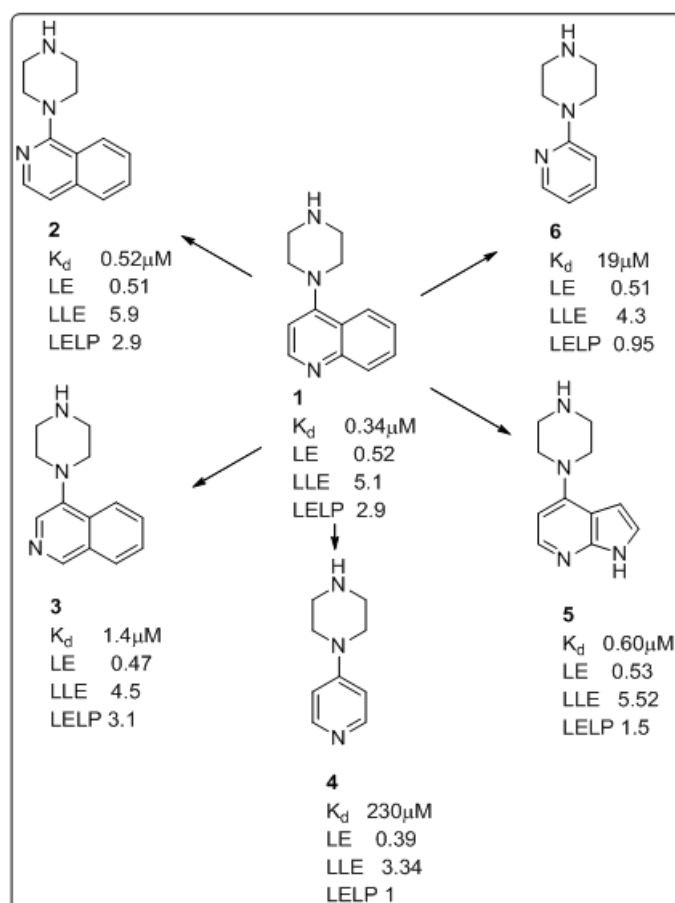


FIGURE 3.7: Scaffold Optimisation

3.6 Results and Discussion

3.6.1 Core Optimisation

First, core optimisation was undertaken (Figure 3.7). Starting with the undecorated quinoline core (1) the isoquinolines (2 and 3), the azaindole (5), and the pyridine scaffolds (4 and 6) were prepared. Comparing the K_D values for the compounds it can be seen that compounds 1, 2 and 5 have the highest K_D values of 0.34 μM , 0.52 μM and 0.60 μM respectively (Figure 3.7), with very similar ligand efficiencies. To investigate the influence of the second fused aromatic ring the pyridine scaffolds (4 and 6) were synthesised. In the case where the pyridyl nitrogen was para to the piperazine (compound 4), the second aromatic ring improved the binding affinity and ligand efficiency, suggesting that it is involved in an important interaction here. However, when the nitrogen was ortho to the piperazine ring (6) the ligand efficiency was more similar to compound 1. Interestingly looking at the Ligand Lipophilicity Efficiency (LLE) values, the presence of the second aromatic ring in the quinolones appears not be driven solely by lipophilic interactions. It was decided to pursue the quinoline scaffold, which had the best ligand efficiency and a good LLE (higher than 5). This scaffold also had more options for vector optimisation and there was commercial availability of substituted quinolines. To optimise the vectors, we substituted as many positions as possible around the quinoline ring with a methyl group. Additional substituents, such as F, Cl and CF_3 were used as well to provide more information. Attempts to substitute hydroxyl groups proved synthetically challenging in most vector directions. Table 3.8 presents the activity for the derivatives with vectors located at different positions around the quinoline ring. We can conclude that when the substituent at R^1 is Me or CF_3 (7 and 10), a 5-10-fold increase in binding affinity was observed when compared to unsubstituted quinoline (1). In the case of compound 7 there was an increase in ligand efficiency and LLE, whereas with compound 10 the increase in binding affinity was countered by the increase in number of heavy atoms, leading to a slight drop in ligand efficiency. There was also a drop in LLE of compound 10, suggesting that some of the activity may be driven by increased lipophilicity of the ligand. The increase in binding affinity might suggest a lipophilic pocket in the active site. Compounds 8 and 9 are the isoquinoline analogues of the scaffold. When substituting quinoline with isoquinoline, affinity and LE were improved. This additional binding was not driven primarily by non-specific lipophilic interactions since the LLE increases. Compounds 8 and 9 both have a substituent in the R^1 position and a change of the position of the nitrogen. We can see a small rise in binding affinity when R^4 is substituted with small and hydrophilic fluorine as seen in compound 9. Compound 9 is the most potent compound from the series with a K_D of 0.0024 μM and LE of 0.58. Comparing compounds 10 and 11, we can observe that Me in R^2 caused a very

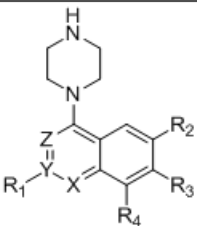
										
Cmpd	X	Y	Z	R ¹	R ²	R ³	R ⁴	K _D (μM)	LE	LLE
1	N	C	CH	H	H	H	H	0.34	0.52	5.1
7	N	C	CH	Me	H	H	H	0.039	0.56	5.5
8	CH	C	N	Cl	H	H	H	0.010	0.61	6.1
9	CH	C	N	Cl	H	F	H	0.0024	0.62	5.9
10	N	C	CH	CF ₃	H	H	H	0.073	0.46	4.6
11	N	C	CH	CF ₃	Me	H	H	0.021	0.47	4.9
12	N	C	CH	H	H	H	Me	3.8	0.41	3.6
13	N	C	CH	H	H	CF ₃	H	3.5	0.35	3.0
14	N	C	CH	Me	H	CF ₃	H	0.77	0.37	3.2
15	N	C	CH	OH	H	H	H	0.41	0.49	4.7
16	N	C	CH	CF ₃	MeO	H	H	0.30	0.40	3.6

FIGURE 3.8: The binding affinity and ligand efficiency of analogues with different vectors.

small increase in affinity and ligand efficiency. However, a methoxy (MeO) in the R² position caused a drop in both affinity and ligand efficiency (compound 16). Therefore, if there is any steric space available around the R² position, it is probably limited. When R³ is substituted with CF₃, a drop in affinity and ligand efficiency was observed (13 and 14), suggesting that there is little room for substituents at this position. Likewise, when R⁴ is Me (12), we observed a drop in affinity and ligand efficiency, suggesting that substituents at this position are not tolerated.

When R¹ is OH, as in compound 15, the compound showed a similar binding affinity and ligand efficiency to compound 1 and reduced affinity and ligand efficiency compared to compounds 7 and 10. In compound 15, the hydroxyl may of course be in the alternate “pyridone” tautomeric form. We were interested to understand the effect of the R¹ substituent on the pK_a of the quinoline nitrogen and the effect, if any, this would have on activity. The pK_a values were calculated using Jaguar pK_a [243]. The data is shown in Figure 3.9. This assumes that the compounds are planar and the piperazine nitrogen

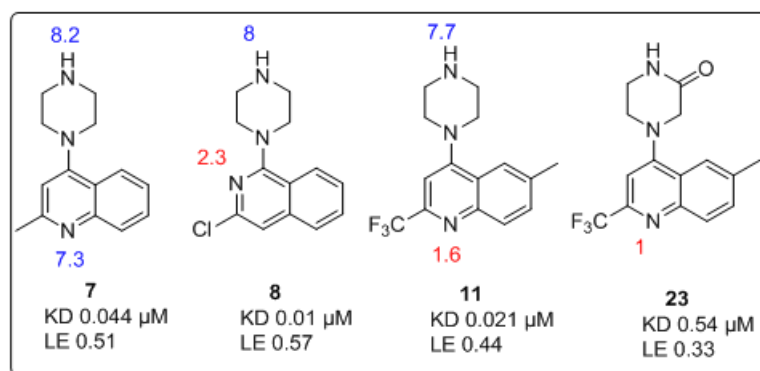


FIGURE 3.9: Calculated pKa values of key functional groups.

is conjugated with the quinoline ring. Docking (see later) suggests that the piperazine ring may be slightly twisted out of plane, which probably reduces the basicity of the quinolone nitrogens. However, it can be concluded that the pKa of the nitrogen in the quinoline ring does not have influence on the binding affinity. Comparing compounds 7 and 11, there is no significant difference in affinity, despite a change in pKa from 7.3 to 1.6.

Table 3.10 shows data when the piperazine moiety was modified. Some interesting points can be deduced here. Comparing compounds 17 and 18, there is a 100-fold drop in binding affinity on methylating the piperazine NH. Compound 18 will be predominantly protonated at physiological pH and so potentially is an H-bond donor. Possibly either the steric bulk of the methyl substituent provides a steric clash with the receptor or the presumed equatorial orientation of the methyl on the piperazine nitrogen prevents H-bond formation (with Asn312^{7,39} – see docking studies later). The docking studies results suggest the latter, where there is a hydrogen bond predicted between the equatorial hydrogen and Asp113^{3,32} for compounds 10 and 11. Comparing 17 with 19 and 20, both the hydrogen bond donor and positive charge have been removed and there is an even larger drop in binding affinity. In lactam 23, the H-bond is retained, but the basic centre removed. Here there was a 25-fold loss in binding affinity compared to compound 17. All this data suggests that both the H-bond donor and basic group are required for strong binding. When R⁶ and R⁷ were substituted with Me (compound 21) both affinity and LE decreased dramatically - this might suggest steric clashes. Compound 22 lost both LE and affinity, with affinity decreasing around 10-fold; this supports the hypothesis that there is a limited steric space around the piperazine part of the molecule which can tolerate only one methyl. We also tested some of the quinoline compounds with dimethyl amino pyrrolidine at position 4 rather than the favoured piperazine (Table 3.11). We can see that this substitution was beneficial for both ligand efficiency and affinity. Compound 25 was one of the most potent compounds from the series with KD 0.005 μM . Whilst this compound does not have an explicit hydrogen to form an H-bond

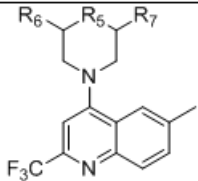
						
Cmpd	R ⁵	R ⁶	R ⁷	K _D (μM)	LE	LLE
17	NH	H	H	0.021	0.47	4.9
18	NMe	H	H	1.9	0.37	2.5
19	CF ₂	H	H	212	0.21	-0.53
20	O	H	H	No binding	0	0
21	NH	Me	Me	540	0.19	-0.53
22	NH	Me	H	0.29	0.39	3.2
23	NH	O	H	0.54	0.37	3.9

FIGURE 3.10: Modifications to the piperazine ring.

on the dimethylamino group, it is presumably protonated at physiological pH, forming a H-bond donor, predicted to H-bond to Asp113^{3,32}. Further the N-methyl makes a lipophilic interaction with Trp1093.28 (explained in the docking section).

3.6.2 Docking

Subsequent to the chemistry work, a docking study was undertaken in order to correlate the predicted binding modes with the SAR observed in the β_2 AR SPR assay. Given the number of molecules and the high fidelity of pose prediction that we were able to achieve in the past [222], docking can be considered the method of choice in this context. There is a crystal structure of the β_2 AR (PDB 2RH1) with carazolol bound [221, 226] (Figure 3.12, Numbering of residues is according to Ballesteros-Weinstein [244]). Carazolol is a high affinity inverse agonist of the β -adrenergic receptors and binds into the orthosteric pocket of the receptor, which is a binding pocket for the derivatives described in this paper, as shown by the pharmacology assays described in Aristotelous *et al* [120]. The compounds were docked into the classical, orthosteric pocket of β_2 AR using Fast Exhaustive Docking (FRED) [58, 81] (see Figure 3.12 for a map of the β_2 AR orthosteric pocket). Test calculations were also performed with the β_2 AR in an active conformation (PDB 4LDL) and yielded differing poses, which is consistent with expectation for inverse agonists or antagonists. The poses of the compounds were scored and the best-ranking poses compared to the findings from SPR assays to rationalise and explain the binding

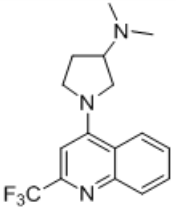
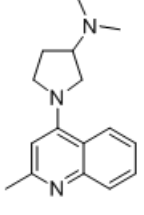
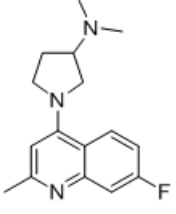
Cmpd	Structure	K _D (μM)	LE	LLE
24		0.55	0.37	2.5
25		0.0059	0.56	5.0
26		0.63	0.40	2.3

FIGURE 3.11: Modifications to the piperazine.

affinities obtained. Using this information, it was possible to compare our “virtual image” of the binding site obtained by SPR experiments with the protein structure. The two binding site images overlapped to a high degree. The docking campaign also suggested possible future work which should further improve the affinity of the fragments.

Analysis of the top ranked poses of the compounds revealed a number of key interactions with the binding site. Compound 1 (the undecorated 4-piperazine quinoline scaffold) was used as a reference compound, when analysing the docking (Figure 3.13 and 3.16). Key compounds are discussed in the main text, while the remaining molecules are examined in the SI.

- The basic nitrogen of the piperazine is in close proximity to the side chain carboxylic acid of Asp113^{3.32} (2.6Å) which will give a charge-assisted H-bond (a salt-bridge).
- The rest of the piperazine fits into a narrow pocket forming lipophilic interactions with the side chains of Trp109^{3.82}, Ile169^{4.61}, Phe193^{5.32} and Tyr308^{7.35}.

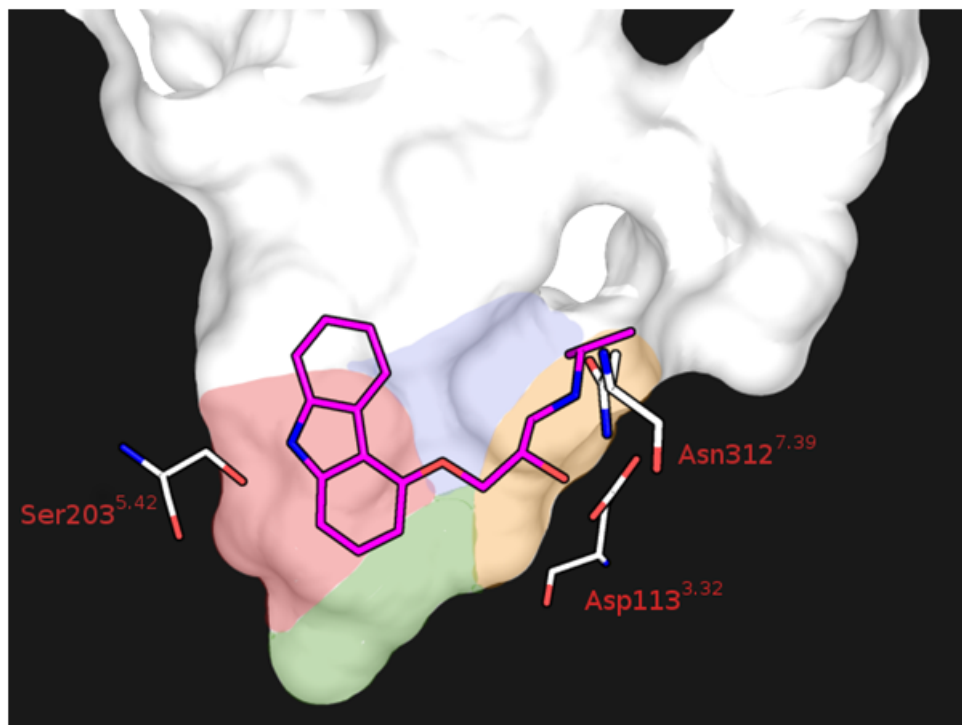


FIGURE 3.12: Coloured map of the binding pocket illustrating carazolol (magenta carbons) bound to the β_2 AR (PDB ID 2RH1). The yellow zone is lined by residues Asp113^{3.32} and Asn312^{7.39}; the green zone is bordered by Ile121^{3.40}, Val117^{3.36}, Phe208^{5.47}, and Phe282^{6.44}; the red one is formed by Ser203^{5.42}, Ser204^{5.43}, and Ser207^{5.46}; and finally, the blue zone is lined by Thr195^{5.34}, Tyr199^{5.38}, Phe290^{6.52} and Asn293^{6.55}

- The quinoline moiety binds to a cavity/ surface bounded by Val114^{3.33}, Thr195^{5.34}, Tyr199^{5.38}, Phe289^{6.51}, Phe290^{6.52} and Asn293^{6.55} (blue zone in Figure 3.12).
- The quinoline nitrogen could potentially form a H-bond interaction with Ser207^{5.46} (Figure 3.13 and 3.16). When docked as the protonated form of the quinoline, compound 1 had a very similar binding mode to that of the un-protonated form (when it acted as an H-bond acceptor from Ser207^{5.46}). In general the binding of the compounds did not seem to correlate with the pKa of the quinolone nitrogen. In addition, the isoquinoline analogues (8 and 9) both are very potent, but cannot make this interaction. Therefore we can conclude that this interaction was not particularly important, perhaps due to the penalty for desolvation of the quinoline nitrogen on binding.
- An examination around the quinoline binding site showed that the protein binding site is mostly hydrophobic. At the bottom of the pocket there are polar residues Ser203^{4.42}, Ser204^{5.43} and Ser207^{5.46} (red zone in Figure 3.12). On the front there is a small lipophilic pocket lined by residues Ile121^{3.40}, Val117^{3.36}, Phe208^{5.47} and Phe290^{6.52} (green zone in Figure 3.12). Vector R¹ points in this direction and when the substituent is lipophilic we observed an increase in affinity.

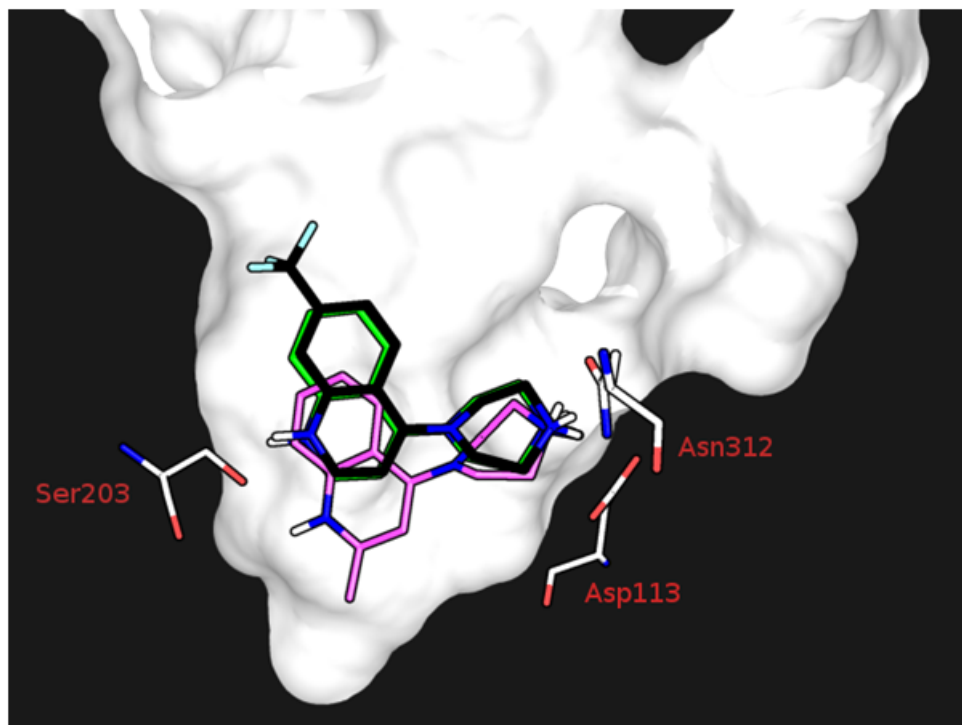


FIGURE 3.13: Undecorated 4-piperazine quinoline scaffold (compound 1, green carbons), R³-decorated scaffold (compound 13, black carbons) and R¹-decorated scaffold (compound 7, pink carbons) docked into the binding site of β_2 AR. Compounds 7, 10 and 11 were more potent than compound 1. On docking these into the binding site, the R¹ substituent in compound 7 appeared to have a lipophilic interaction with the active site as does the hydrophobic substituent at R² for compound 11. Interestingly the molecule is predicted to flip around the piperazine axis between compound 1 and compound 10 (Figure 3.17) and 11. This flip could be caused by the CF₃ group forming halogen interactions with Asn293^{6.55}.

A map showing the different regions of the orthosteric binding site is shown in Figure 3.12.

Figures 3.13 and 3.18 also show the predicted binding pose of compound 13, which differs from compound 1 in that there is a CF₃ group at the R³ position. In this pose, due to the steric constraints the CF₃ group does not occupy a hydrophobic part of the active site, but rather points towards the polar atoms of Asn293^{6.55}. This may explain the SPR data in which affinity drops by around 10-fold when compared to the un-substituted quinoline (compound 1). When looking at the docking pose, the general orientation of the scaffold is retained compared to compound 1 (Figure 3.13), and there's enough space around R¹ to be filled with some other substituents, preferably hydrophobic/ lipophilic Me/CF₃.

Almost all of the synthesised compounds occupy the same space in the pocket, and the H-bond formed between piperazine and Asp113^{3.32} anchors the compounds. When piperazine is substituted with di-methyl-aminopyrrolidine (25), the compound showed

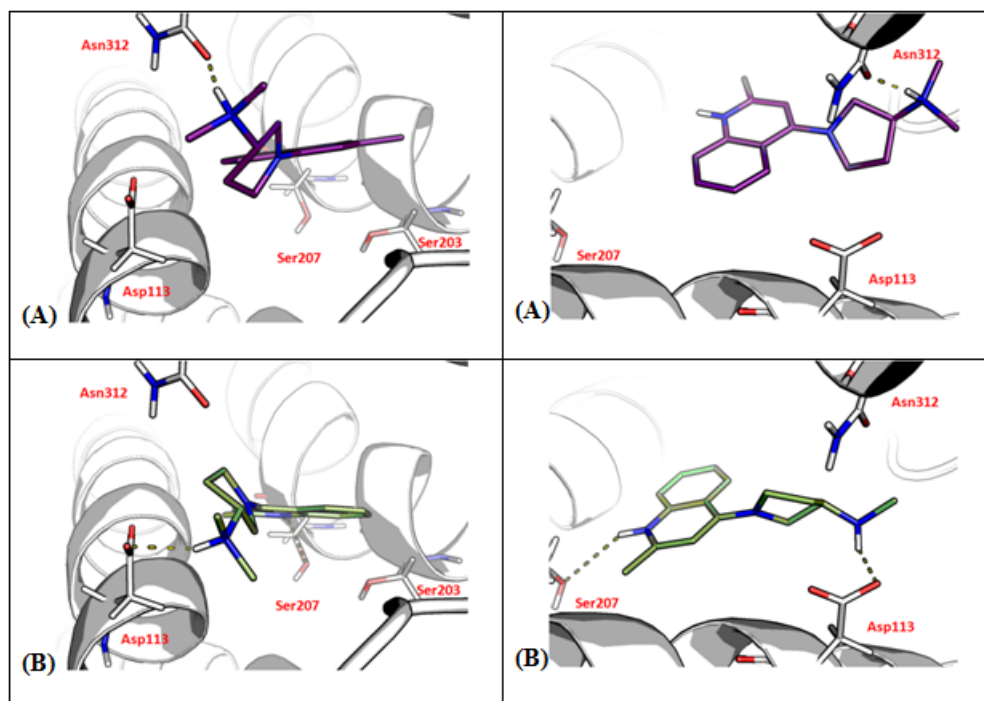


FIGURE 3.14: Compound 25 docked into the β_2 AR structure. (A) The R-enantiomer; (B) the S-enantiomer.

an improved affinity in the SPR studies. The dimethylamino group is presumably protonated at physiological pH and can still form a salt-bridge with Asp113^{3,32}. There are two enantiomers for compound 25. When docked each of the enantiomers has a different binding mode. The S-enantiomer has a binding mode, where the methyl at R¹ fits in the hydrophobic part of the with Asp113^{3,32} (Figure 3.14 (B)) pocket and the NH of pyrrolidine forms a H-bond. There may also be a lipophilic interaction between one of the methyl groups in the pyrrolidine part of the molecule and Trp109^{3,28}. A hydrogen bond interaction was observed between the nitrogen from the quinoline ring and Ser203^{5,42}. This is a favourable set of interactions. In contrast, the R-enantiomer has the quinoline ring flipped, causing loss of the critical H-bond with Asp113^{3,32} and losing the lipophilic interactions with the binding pocket (Figure 3.14 (A)).

3.7 Conclusion

In this paper we have reported the optimisation of fragment hits of the β_2 AR. The basis of this optimisation was a decision tree for core and vector optimisation. The rationale for this tree was that even small changes in either the core or the vectors can have a substantial impact on the activity of a molecule, in particular a fragment. This would then give valuable information on how to optimise a fragment in the absence of structural information. We were able to demonstrate that this was the case. By careful

analysis of the ligand efficiency, this should show whether a substituent was favourable or non-favourable. In addition, the use of the lipophilic ligand efficiency is helpful in determining whether the changes are driven by non-specific lipophilic interactions.

In this case, we had the advantage of a starting point with a good ligand efficiency. These fragments were successfully optimised using this process in the absence of structure. By careful choice of small substituents, we were able to rapidly understand potential vector for optimisation of the hits. In this case, there was information available on the structure of the receptor and a ligand co-crystallised with the receptor. By docking experiments with the ligands prepared in this project, we were then able to rationalise the results, which gives added confidence to our decision tree. It can be concluded that our decision tree approach has helped the optimisation process and it can be used to grow the fragments further, albeit in this case the starting fragments had an unusually high potency.

3.8 Acknowledgements

The authors would like to acknowledge the Wellcome Trust for support (SG, WT097818/Z/11A and IHG WT100476). This work is part-funded by the MSD Scottish Life Sciences fund (TA). As part of an ongoing contribution to Scottish life sciences, Merck Sharp and Dohme (MSD) Ltd (known as Merck and Co., Inc., Kenilworth, NJ, USA, in the United States and Canada), has given substantial monetary funding to the Scottish Funding Council (SFC) for distribution via the Scottish Universities Life Sciences Alliance (SULSA) to develop and deliver a high quality drug discovery research and training programme. All aspects of the programme have been geared towards attaining the highest value in terms of scientific discovery, training and impact. The opinions expressed in this research are those of the authors and do not necessarily represent those of MSD, nor its affiliates. PK thanks the German Research Foundation DFG for Emmy Noether grant KO4095/1-1. PK and IG participate in COST Action CM1207 "GLISTEN" and parts of the research has been done during a Short-Term Scientific Mission of SG to the lab of PK. We would like to thank Professor R. Lefkowitz for supply of protein.

3.9 Supporting Information

3.9.1 Methods

3.9.1.1 Chemistry

General Methods Chemicals and anhydrous solvents were purchased from commercial sources and were used without further purification. ^1H NMR spectra were recorded on Bruker Avance DPX 500 spectrometer. Chemical shifts (δ) are expressed in ppm. Signal splitting patterns are describe as singlet (s), broad singlet (bs), doublet (d), triplet (t), quartet (q), multiplet (m) or combinations thereof. LC-MS analyses were performed with either an Agilent HPLC 1100 series connected to a Bruker DaltonicsMicroTOF, or an Agilent Technologies 1200 series HPLC connected to an Agilent Technologies quadrupole LC/MS, both instruments were connected to an Agilent diode array detector.

General Procedure All compounds were made using the following general procedure. A mixture of 4-chloroquinoline (0.30 mmol), amine (0.60 mmol) and K_2CO_3 (0.75 mmol) in DMF (5 mL) was stirred at 120°C for 16 h. After monitoring the end of the reaction by LC-MS, the mixture was concentrated in vacuo and the residue was partitioned between EtOAc (20mL) and brine (20 mL) twice. The combined organic layers were dried over MgSO_4 and the solvent was removed in vacuo to afford the compound. LCMS data were obtained using electrospray and showed final compounds had purity higher than 95%.

3.9.1.2 SPR methodology

A human β_2 adrenoceptor construct containing a FLAG tag at the N-terminus and histidine 10 (His-10) tag at the C-terminus was generated for baculovirus expression in Sf9 cells. The receptor was solubilized and purified as described before [245]. SPR assay for kinetic characterisation validating activity of the β_2 adrenoceptor sensor surface was described previously [120]. The same method was applied in this study. Briefly, the β_2 adrenoceptor was captured via His-10 tag on NTA sensor chip in capture buffer consisting of 50 mM Hepes pH 7.4, 150 mM NaCl, 50 uM EDTA, 3% DMSO, 0.01% MNG (lauryl maltose neopentyl glycol) obtaining capture levels 10,000-11,000 RU. Biacore T100 and T200 were used for all experiments, and in order to increase stability of the β_2 adrenoceptor on the surface all experiments were conducted at 10°C to increase stability of the receptor on the surface. Solid compounds were solubilized in 100% DMSO and

then diluted into running buffer to a final DMSO concentration 3%. For kinetic analysis compounds were screened at 30 – 60 s for association and 60 – 600 s for dissociation at concentrations adjusted according to the affinities (2-fold or 3-fold dilutions between 300 μ M and 0.23 nM – total highest and lowest concentration, the selection of highest and lowest concentration was adjusted for each compound separately to obtain suitable concentrations for kinetic fit) at flow rate 30 μ L/min. Scrubber 2 software (BioLogic Software, Australia) was used to fit kinetic data. 1:1 model including mass-transport limitation was used, affinity was calculated as ratio of calculated parameters for off-rate/on-rate (k_a/k_d).

3.9.1.3 Docking preparation

The docking calculations were performed with the β_2 AR inactive structure in complex with carazolol (PDB: 2RH1) [220, 221] using FRED [58, 80, 81] and up to 10 poses were generated for each compound. All ligand, solvent and lipid molecules were removed. All compounds studied through docking were subject to conformer generation using OMEGA [57], with an RMSD of 0.1 Å. The protonation states were defined using QUACPAC [246]. Binding modes were chosen based on their interaction with Asp113^{3.32} and lack of interaction violations, e.g. non-interacting hydrogen-bond donors, charge mismatches and unlikely torsion angles.

3.9.2 Results

3.9.2.1 SPR

3.9.2.2 Docking

The quinoline nitrogen is likely to be a much poorer H-bond acceptor in compound 10, compared to compound 1, due to the electron withdrawing characteristics of the CF₃ group, weakening potential interactions with Asn293. Compounds 10 and 11 differ only by one methyl and are predicted to bind in the same way. Compound 11 gave a marginally better affinity (20nM against 70nM) and also is predicted to have higher lipophilicity ($\text{clogP} = 3.4$ against 2.9). This extra binding energy may well be driven by lipophilic interactions of methyl with the hydrophobic part of the pocket. The higher LLE for compound 11 suggests this lipophilic interaction is specific.

In the docking pose of compound 12 (Figure 3.19) we can observe the influence of the Me substituents at position 8 (R4). In SPR studies we have observed a significant drop in both affinity (10 fold) and ligand efficiency compared to the undecorated quinoline core

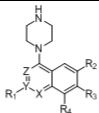
											
Cmpd	X	Y	Z	R1	R2	R3	R4	K _a	K _d	K _D (μM)	LE
1	N	C	CH	H	H	H	H	1.04 (±0.01) × 10 ⁶	0.357(±0.001)	0.3445 (±0.00006)	0.54
7	N	C	CH	Me	H	H	H	1.33 (±0.01) × 10 ⁶	0.0527(±0.0004)	0.0395 (±0.0002)	0.59
8	CH	C	N	Cl	H	H	H	1.840 (±0.009) × 10 ⁶	0.0238(±0.0001)	0.01	0.64
9	CH	C	N	Cl	H	F	H	2.972 (±0.007) × 10 ⁶	0.00710(±0.00001)	0.002388(±0.000002)	0.65
10	N	C	CH	CF ₃	H	H	H	3.771 (±0.007) × 10 ⁵	0.02745(±0.00005)	0.07280(±0.00008)	0.46
11	N	C	CH	CF ₃	Me	H	H	8.04 (±0.02) × 10 ⁵	0.01698(±0.00005)	0.02113 (±0.00009)	0.43
12	N	C	CH	H	H	H	Me	NA	NA	3.8	0.44
13	N	C	CH	H	H	CF ₃	H	9.95 (±0.08) × 10 ⁴	0.346(±0.003)	3.473 (±0.004)	0.37
14	N	H	H	Me	H	CF ₃	H	NA	NA	0.77	0.40
15	N	H	H	OH	H	H	H	7.09(±0.02) × 10 ⁴	0.02932(±0.00006)	0.4135(±0.0006)	0.51
16	N	H	H	CF ₃	MeO	H	H	2.27(±0.01) × 10 ⁵	0.0693(±0.0005)	0.305 (±0.001)	0.39

FIGURE 3.15: Data for key scaffold hops.

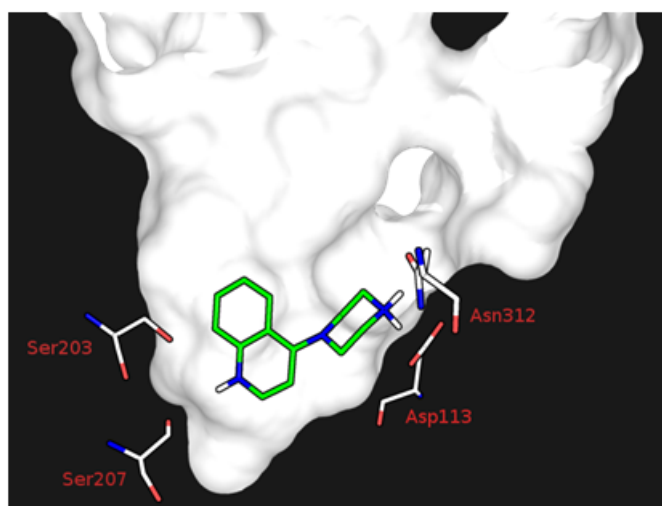


FIGURE 3.16: Undecorated 4-piperazine quinoline scaffold (compound 1), docked into the binding site of β 2AR.

(compound 1). Such a change in affinity can be explained by the docking model. The Me substituent at position R4 prevents the quinoline from sliding deep into the pocket, and changes the orientation of the ring in the pocket to the less favourable orientation. Loss of potency might be due to the lack of interaction with the bottom hydrophobic part of the binding pocket. The movement of the quinoline core probably reduces the lipophilic interactions of the phenyl moiety with Tyr199 and Phe193. Potential stacking interactions with Asn293 are also lost.

The methyl substitutions on the piperazine ring (compound 21, figure 3.20) decreased the binding affinity to the receptor by more than 1000-fold. In the docking prediction

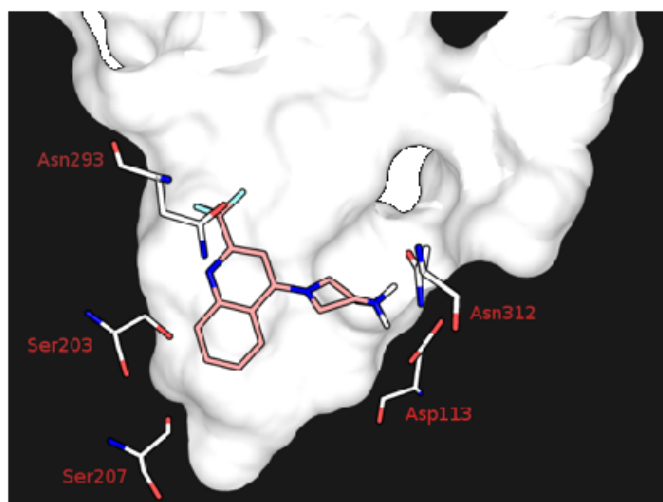


FIGURE 3.17: Compound 10 docked into the β_2 AR structure. Piperazine forms an H-bond with Asp113, through the equatorial hydrogen.

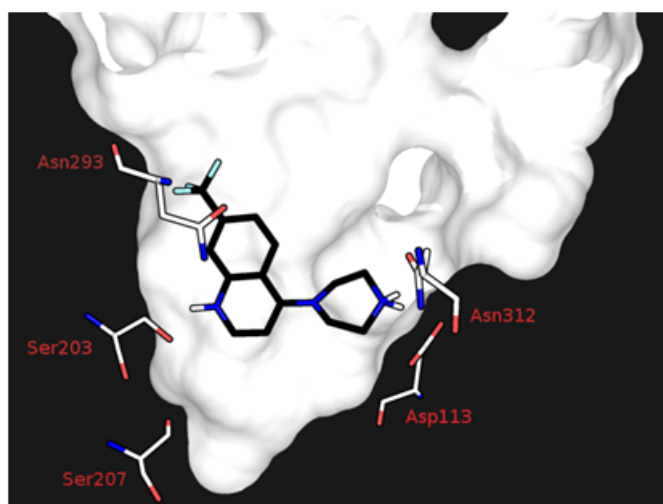


FIGURE 3.18: Compound 13 docked into the β_2 AR structure.

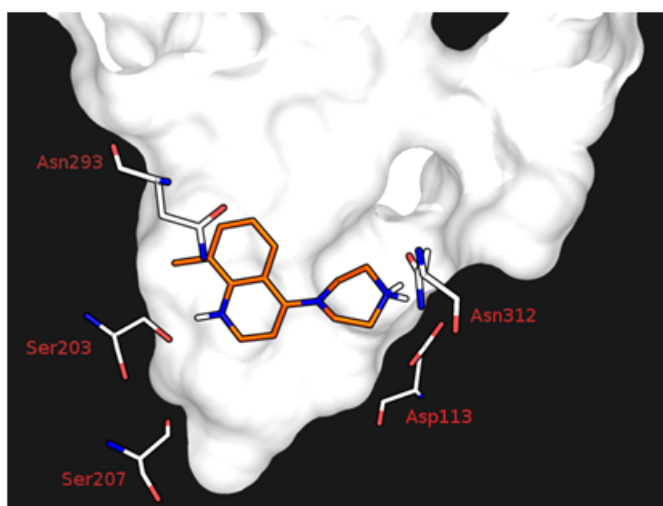


FIGURE 3.19: Compound 12 docked into the β_2 AR structure

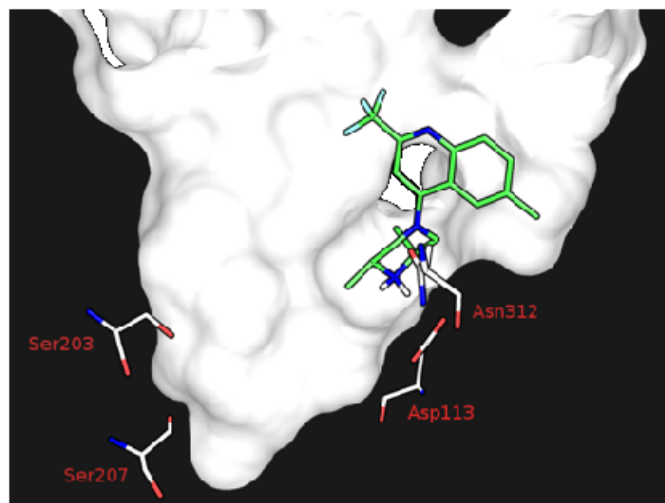


FIGURE 3.20: Compound 21 docked into the β_2 AR structure

the disubstituted piperazine is too wide to slide through the narrowest part of the target, preventing the quinoline moiety molecule from entering into the active site.

"Growing strong."

House Tyrell, *A Song Of Ice And Fire*

The chapter 4 is an article about to be submitted. The submission is bound to a companion paper of our biologist collaborator, where they will describe their new assay. The authors list is the following (by contribution order): Chevillard F., Rimmer H., Betti C., Pardon E., Ballet S., Steyaert J., Diederich W. E., and Kolb P. I was responsible of the overall growing strategy, creation of the PINGUI workflow, selection of the molecules to synthesize, online implementation of the PINGUI toolbox (within SCUBIDOO) and the pictures generation. Helena Rimmer synthesized three compounds (K010) and Cecilia Betti synthesized five compounds (K011). Helena is second author because she worked on our initial hit (Z32501319) identified by Peter Kolb. Els Pardon was responsible of the assays.

Chapter 4

PINGUI: Binding-site compatible growing applied to the design of β_2 -adrenergic receptor ligands.

4.1 Abstract

Fragment-based drug discovery is intimately linked to fragment extension approaches, such as growing, merging or linking. These approaches can be accelerated using softwares or computational workflows for *de novo* design. Although computers allow for the facile generation of millions of suggestions, this often comes at a price: uncertain synthetic feasibility of the generated compounds, potentially leading to a dead end in an optimization process.

In this study we computationally extended, chemically synthesized and experimentally assayed new ligands for the β_2 -adrenergic receptor (β_2 AR) by growing fragment-sized ligands. Our approach is based on the assumption that each individual building block to be added should engage in favorable interactions with the protein on its own. In order to address the synthetic tractability issue, our *in silico* workflow aims at derivatized products based on robust organic reactions. Hence, growing is guided by both the protein structure and synthetic feasibility.

The study started from the predicted binding mode of five fragments. All five fragments were predicted to bind within the orthosteric site of the β_2 AR. We suggested a total of eight diverse extensions aiming to fill the secondary binding pocket (SBP) adjacent to the orthosteric site. The eight compounds were successfully synthesized and further

assays showed that four products had an improved affinity compared to their respective initial fragment.

The described workflow can improve early fragment-based drug discovery projects, especially in the realm of fragment growing strategies because it suggests extensions that are very likely to be attachable, making it a useful creative tool for medicinal chemists during structure-activity relationship (SAR) studies. We made the bulk of the computer-aided approaches that were used in this study freely accessible online. This toolbox called PINGUI aims at assisting fragment-based growing projects towards new products which offer a high likelihood of synthetic tractability.

4.2 Introduction

In the past few years, fragment-based drug discovery (FBDD) has continuously gained in popularity and has become a dominant approach in order to explore novel chemical entities [6, 10]. The year 2011 witnessed an important success for this field, with the first drug (Vemurafenib), originating from a hit found in a fragment-based screening, reaching FDA approval [143, 247].

Fragments are commonly regarded as attractive alternatives compared to standard compounds contained in the libraries used for HTS campaigns. They are small molecules ($MW < 300$ Da) and usually offer better solubility, making them favorable in terms of ADMET (absorption, distribution, metabolism, excretion and toxicity) properties [22]. Moreover, it's easier to cover chemical space with fragment-sized molecules: screening a set of 1'000 fragments is claimed to be equivalent to probing the chemical space of 1'000'000 drug-like molecules [17]. Because of their simplicity and few polar interaction points, they also have a higher chance to bind to a given protein with optimal interactions. However, because they are smaller, fragments usually bind with low affinity compared to drug-like molecules, resulting in the need for screening at high concentrations. Yet, those that bind do so with high ligand efficiency (LE), defined as the ratio of Gibbs free energy (ΔG) and the number of heavy atoms [14]. Fragments are thus often considered advantageous starting points for exploratory synthesis campaigns.

Several biophysical techniques can deal with the low affinity and therefore high concentration of fragments and allow detection of binding to a protein. Among them, surface plasmon resonance (SPR), isothermal calorimetry (ITC), thermal shift assays and mass spectrometry (MS) have been successfully implemented in fragment-based drug discovery pipelines [110, 111]. However, since the interactions remain weak, structural methods such as X-ray crystallography[112] and NMR[113] might be preferable, as

they lead to fewer false positives during screening campaigns [17]. These two structural techniques can in principle determine the precise binding modes of a fragment within a protein target, thereby also laying the foundation for subsequent structure-based design approaches.

When no experimental data is available, the binding mode of a fragment within an active site can be predicted using *in silico* methods, most notably docking. It has to be noted that, because of their propensity to bind in more than one mode, fragment binding mode prediction remains challenging. Moreover, most of the docking programs were parametrized and optimized for drug-like molecules, which might lead to a relative imbalance of the individual terms in a scoring function [22]. Consequently, several programs have been developed more specifically for fragment docking, including MCSS [103], SuperStar [104, 105] and SEED [77, 78].

Once the binding mode of a fragment has been determined, either experimentally or computationally, it can be used as starting point for an FBDD project, aiming at expanding the fragment into a bigger molecule, while concomitantly increasing its potency.

Three main approaches are conceivable in order to extend fragments: merging, linking and growing (Figure 2.5). *Merging* concatenates two fragments that contain a common portion known to bind at the same position in a given target. The *linking* strategy describes the process of joining two non-competitive fragments (i.e. fragments that bind in two different regions of the binding site). *Growing* is based on only one fragment and aims at extending it within the binding site, looking for additional interactions that could improve affinity or selectivity. *Growing* is the most frequently used approach because it is considered more straightforward and successful than the other techniques [177, 178].

Growing approaches have extensively been implemented in computer softwares, including LUDI [89], SPROUT [185], CONCERTS [186], ReCore [187], Caveat [188], BREED [189], GANDI [69] and BROOD [190]. However, one of the main challenges of these softwares is to assure the synthetic tractability of the designed molecules. Several rules can be implemented in order to assess whether or not a compound is synthetically feasible, among the most frequently used are BRICS [191], RECAP [192] or an estimation of the so called synthetic accessibility score [193]. These approaches are based on knowledge of chemical reactions in order to analyze molecules.

Proceeding the other way around is also possible: attaching two available building blocks using a compatible chemical reaction. Hartenfeller *et. al.* developed a library of 58 robust organic reactions [1, 2], based on often-used chemical reactions in the pharmaceutical field [248]. These reactions are at the core of our PINGUI (Python In silico de Novo Growing Utilities) toolbox and define the first step of this study: creating a focused

library consisting only of building blocks that are compatible with the core fragment for further synthesis.

Secondly, our approach for fragment growing is based on the assumption that each individual building block should engage in favorable interactions with the protein by itself. In the alternative approach, namely pre-attaching extensions to core fragments and re-docking the resulting compounds, slightly unfavorable extensions might be compensated and thus masked. Within the PINGUI workflow we generate *maps* containing a large number of favorable poses for each extension building block. Based on the crystallographic or computed binding mode of the core fragment, building blocks are then chosen such that reactive groups are within geometrically feasible distances and angles, therefore allowing bond formation between the core and an already optimally placed building block to occur. In this way, fragment chemical and positional space is pruned at an early stage and only optimally positioned fragments with favorable interactions are retained. Re-docking of the computationally derived products ensures that the initial assumption (i.e. that each component of a molecule is, when regarded individually, located in an optimal spot) is still valid. A large deviation from the original (predicted) binding mode would violate this assumption and lead to the removal of the product from further consideration. Hence, growing is guided by the protein structure and the synthetic tractability.

In this manuscript, we describe the design of new ligands of the β_2 AR by employing the PINGUI toolbox. Five different fragments were grown and have in common that they are predicted to bind in the orthosteric site with a reaction-compatible reactive feature (herein an amine) near the SBP (Figure 4.1). Therefore, each growth was directed towards the SBP. Experimental results demonstrated that all the eight suggested molecules were successfully synthesized and half of them had a more favorable affinity than the initial fragment.

4.3 Methods

4.3.1 The β_2 AR binding site

The β_2 AR binding site sits buried in the extracellular part of the receptor with only a small fraction accessible to solvent [222]. For the sake of this study, we divided the binding site in two smaller cavities: the orthosteric site and the SBP (Figure 4.1). While most of the orthosteric site is hydrophobic, important polar residues shape its contour: Ser203^{5.42}, Ser204^{5.43} and Ser207^{5.46} in transmembrane helix 5 (TM5) define the bottom of the orthosteric site. These serines are involved in receptor activation [224, 225, 228].



FIGURE 4.1: Sliced surface side-view of the binding site of the β_2 AR in its active conformation (PDB code 2RH1). The orthosteric site is colored in green and the SBP in blue.

Asn312^{7.39} and Tyr316^{7.43} in TM7 and Asp113^{3.32} in TM3 offer strong electrostatic points that interact with the amine moieties of most known ligands. The SBP is defined by a small hydrophobic cavity (Trp109^{3.28}, Phe193^{5.32} and Ile309^{7.36}) and a number of more solvent exposed polar residues (Asp192^{ECL2} and His93^{2.64}) which agonists are known to interact with [224–228].

The difference in the orthosteric site between the inactive and active receptor conformation is manifest mostly in His296^{6.58} [226, 228]. In the active-like receptor conformation (PDB 4LDE), His296^{6.58} is part of a large polar network involving the catechol moiety of the ligand, a water molecule, Tyr308^{7.35}, Ser204^{5.43} and Asn293^{6.55} [226, 227]. This polar network is located close to the orthosteric site, making it smaller in the active conformation (Figure 4.2). In the inactive conformation of the receptor, the aforementioned polar network is not so extended, resulting in an inward shift of His296^{6.58} [226, 228]. This shift results in a larger orthosteric site.

4.3.2 Receptor X-ray structures

Docking calculations were performed with the inactive conformation of the β_2 AR in complex with carazolol (PDB: 2RH1) [220, 221] and a partially active conformation in complex with the ligand BI167107 (PDB: 4LDE) [226]. All ligands, solvent, lipid molecules as well as the T4-lysozyme insertion or the stabilizing nanobody Nb6B9 were removed. The hydrogens were placed and minimized using the HBUILD module in

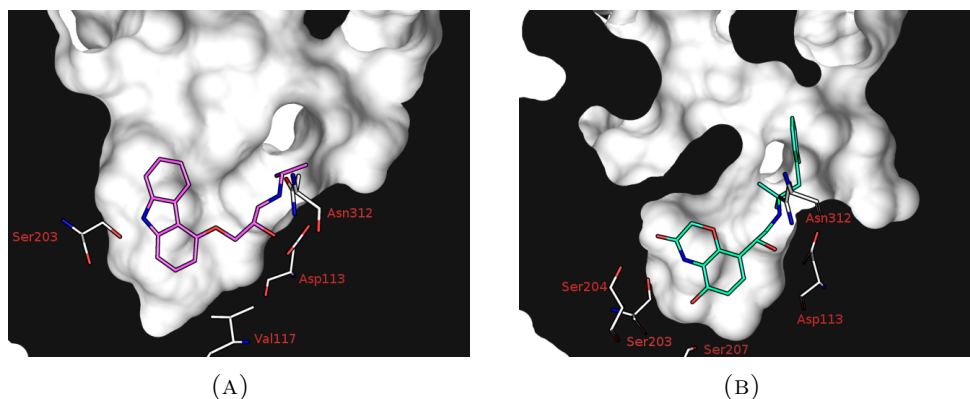


FIGURE 4.2: (a) Carazolol (magenta carbons) in complex with the inactive conformation of the receptor (PDB 2RH1). (b) BI167107 (green carbons) in complex with the active conformation of the receptor (PDB 4LDE).

CHARMM [249]. CHARMM22 [250] atom types and MPEOE [251, 252] partial charges were assigned using the program Witnotp (Novartis Pharma AG, unpublished).

4.3.3 Datasets

4.3.3.1 Core fragments

Five fragment-sized ligands with experimentally determined affinity were used as core fragments (i.e. starting points) in this growing study. Z32501319 (Figure 4.3) was initially discovered through a docking screen against the β_2 AR in an inactive conformation [222] and showed a favorable ligand efficiency (LE) of 0.36. In parallel, four small fragments (Figure 4.4) with high ligand efficiency that had emerged from an experimental screen were also selected.

Fragment Z32501319 displayed higher affinity towards the receptor conformation stabilized in an inactive state by the nanobody fusion. We will refer to molecules that display a similar preference as “inverse agonist candidate” (IAC) to clearly distinguish them from confirmed inverse agonists from experiments based on signal transduction assay.

The remaining four fragments had known a preference for the receptor conformation stabilized in an active state by the nanobody fusion. We will refer to molecules that display a similar preference as “agonist candidate” (AC) to clearly distinguish them from confirmed (partial) agonists from experiments based on signal transduction assay.

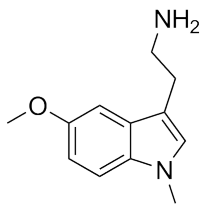


FIGURE 4.3: Z32501319 was chosen as one of the core fragment for our growing strategy.

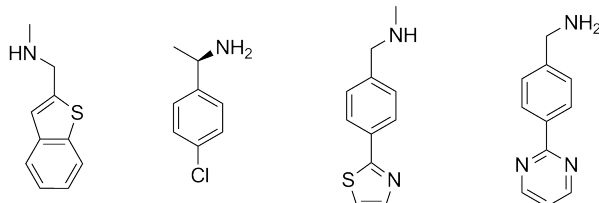


FIGURE 4.4: Four fragments with high LE from an experimental screen were chosen for further growing.

4.3.3.2 Building blocks

The entire *frag now* dataset was downloaded from the ZINC database [253] and contained 504'074 fragments.

4.3.3.3 Surrogates

During the first docking stage, all compatible building blocks (i.e. all fragments amenable to being used as extension of the core fragment) were docked individually in order to find the ones that are likely to fit the SBP and form favorable interactions (i.e. score). However, in their unreacted forms, the building blocks will usually possess atom groups which are not present in the final product. Thus, docking the building block does not reflect its interaction options once attached. In order to mimic the behavior of the final product and minimize the occurrence of unlikely interactions, we converted them to "surrogates". We define these surrogates as the final product without the core fragment. In the specific case presented here, the reaction to be applied to all cores was reductive amination. Surrogates are thus the reactants minus the ketone or aldehyde groups, but with an added amine part (Schemes 1 and 2). In this way, no additional interactions that could influence the placement of the reactant are available. In case of ketones, where substitution by an amine results in the introduction of a chiral center, both stereoisomers for the surrogate were generated using flipper [254]. CHARMM22 atom types and MPEOE partial charges were assigned using the program WITNOTP. Up to 100 conformers were generated for each surrogate using OMEGA [57].

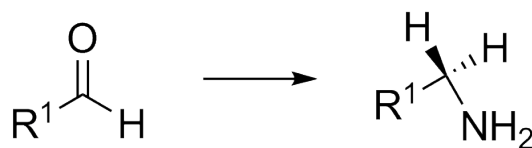


FIGURE 4.5: A surrogate derived from an aldehyde defined by the replacement of the carbonyl by an amine group. R1 = alkyl.

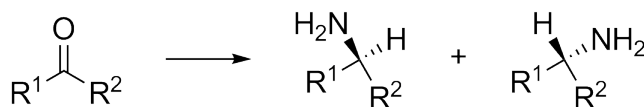


FIGURE 4.6: A surrogate derived from a ketone defined by the replacement of the carbonyl by an amine group. This transformation introduces a chiral center, therefore two surrogates are generated (*R* and *S*). R1 and R2 = alkyl.

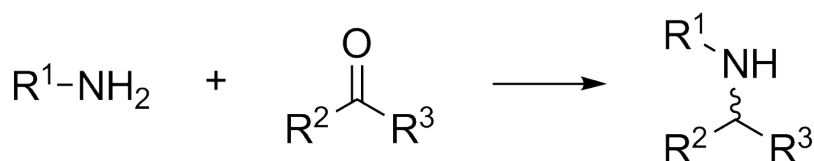


FIGURE 4.7: Reductive amination reaction between a carbonyl and an amine group. R1, R2 = alkyl and R3 = alkyl, H.

4.3.3.4 Products

For each fragment to be grown, the corresponding products were generated from the top 500 surrogates with favorable scores and appropriate geometry using PINGUI (see below). Finally, they were converted into db file format using *dbgen* as incorporated in the DOCK package [255].

4.3.4 Chemical derivatization by reductive amination

All the fragments of this study share a common chemical feature: an amine which will be charged at physiological pH and thus presumably interact with Asp113^{3,32}. This reactive group can also be harnessed to grow the fragment towards the SBP. The most straightforward reaction that can be used with such a functional group is reductive amination (Scheme 3). This reaction involves a primary or secondary amine reacting with an aldehyde or a ketone in order to create a secondary or tertiary amine. The SMARTS code for the reaction was extracted from Hartenfeller et al. [1].

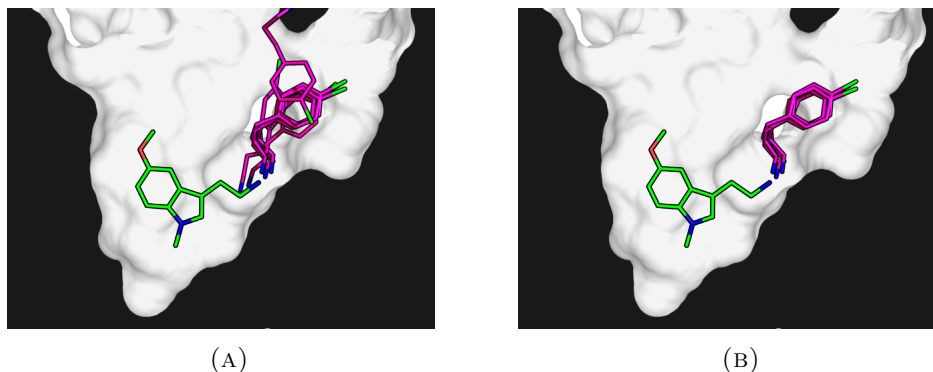


FIGURE 4.8: Illustration of the filtering process. (a) Typical poses (i.e. map) after docking with SEED. (b) The surrogate poses (magenta carbons) were filtered so they do not overlap with the core scaffold (green carbons) and the charged amine is close to the one of the core fragment.

4.3.5 Docking with SEED

4.3.5.1 Sampling and scoring

SEED has been developed to exhaustively place small rigid fragments in the binding sites of proteins. The calculations also take into account the penalty that is incurred upon binding due to the removal of the water shells of the ligand and protein, respectively. For each fragment, $10^6 - 10^7$ individual poses are generated and subsequently clustered based on geometric and energetic criteria. This yields on the order of hundreds of poses, clustered into groups of a maximum of five. These maps were used to select poses with favorable energy and geometrically appropriate orientation with respect to bond formation.

4.3.5.2 Filtering and ranking

The surrogate library was docked with SEED and only those surrogate poses that did not overlap with the core fragment were placed in the SBP as illustrated in Figure 4.8 (a). Then, a filter was applied based on the distance between the amine of the surrogate and the predicted position of the amine of the core fragment (i.e. the docking pose). This cut-off was set to 2 Å. Every pose that passed both criteria was then ranked according to the score calculated by SEED. Only the top 500 surrogates (with no duplicates) were kept for further processing.

4.3.6 Docking with DOCK

4.3.6.1 Core fragments

The four AC fragments and the IAC fragment were placed in the binding site of the receptor active and inactive conformation, respectively, using DOCK 3.6. The predicted binding modes were kept for further growing and visually examined for plausibility (see Results).

4.3.6.2 Products

For each core fragment, all 500 derivative products were placed in the binding site of the β_2 AR using DOCK 3.6. Product conformations were generated using the pipeline described in ZINC [255].

4.3.7 Pose minimization: Szybki

Since DOCK does not evaluate intramolecular energy terms, some docked poses might show unfavorable geometries. In order to ameliorate the clashes of such poses while keeping the overall binding mode found by DOCK, the poses were minimized using the force field including the Poisson-Boltzmann model for solvation as implemented in SZYBKI [87]. All products were then re-ranked according to the score calculated by SZYBKI and kept for subsequent visual inspection.

4.3.8 Workflow of the growing strategy

Figure 4.9 shows a comprehensive scheme of our growing workflow.

4.3.9 Experimental synthesis

4.3.9.1 K011 derivative products

The secondary amines used in this study were prepared through a reductive amination between a carbonyl compound (i.e. the aldehyde or ketone corresponding to the fragment selected) and a primary amine. This reductive amination reaction was performed in two steps. First, the aldehyde and the amine were mixed together until the imine formation was completed and subsequently the imine was reduced to its corresponding secondary amine by means of NaBH₄. After an aqueous work-up, the desired products were isolated

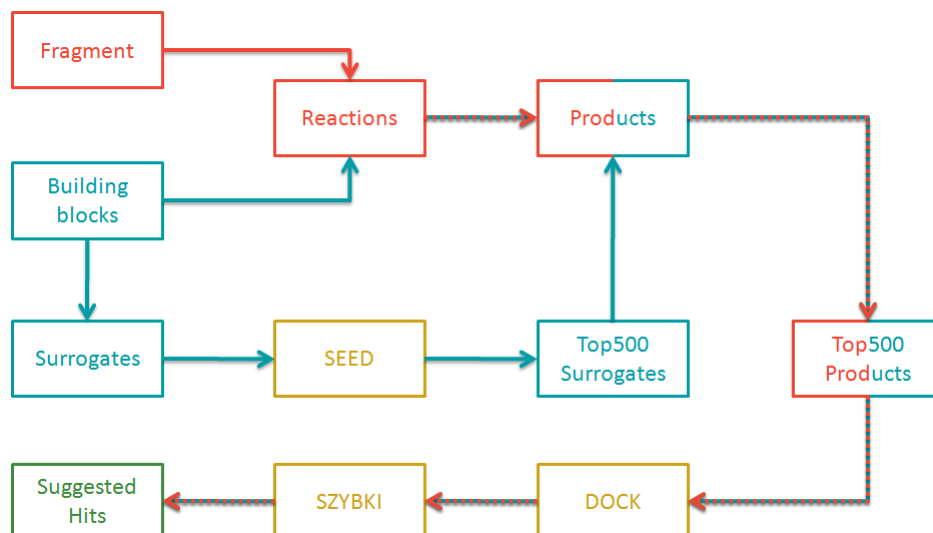


FIGURE 4.9: The PINGUI workflow

by purification on a reverse-phase (RP) semi-preparative HPLC. The structures of the desired products were confirmed by ^1H NMR and mass spectrometry (MS) analysis.

4.3.9.2 Z32501319 derivative products

All commercially available reagents and solvents were used without further purification. Thin layer chromatography was performed on pre-coated plates (silica gel 60 F₂₅₄, Merck). Flash column chromatography was performed on pre-packed columns (PF-30SIHP-JP/ 12G; PF-30SIHP-JP/ 4G; Interchim) using a Büchi separation system. ^1H NMR and ^{13}C NMR spectra were recorded on a Jeol ECA-500 and a Bruker AV II-300 spectrometer. Unless noted otherwise, spectra were recorded at 20°C. Chemical shifts (δ) are given in ppm (parts per million). All NMR spectra were referenced to the residual solvent signal (CDCl_3 : 7.26 ppm [^1H] and 77.16 ppm [^{13}C]). Coupling constants are reported in Hertz (Hz). Mass spectra were recorded on a double-focusing sector field spectrometer type AutoSpec (Micromass). Elemental combustion analyses were recorded on an Elementar vario MICRO instrument.

A solution of the respective aldehyde (1 eq) and amine (1.06 eq) in MeOH (0.1 mol/L) was stirred at room temperature under an Argon atmosphere for 24 h. NaBH_4 (1.6 eq) was added slowly, followed by further stirring for 10 to 15 min. The reaction mixture was quenched with 3M NaOH solution and the product extracted with EtOAc. The organic layer was washed with water, saturated aqueous NaCl, dried over MgSO_4 , filtered, and concentrated *in vacuo*. The crude reaction product was purified by flash column chromatography.

Analytical Data of the Compounds K010FC006, K010FC007 and K010FC008 are provided in SI.

4.3.10 Radio ligand displacement assay

Compounds were examined for their ability to inhibit the binding of [^3H]-dihydroalprenolol ([^3H]-DHA; 2 nM final) to Sf9 membranes expressing $\beta_2\text{AR}$. 5 μg of total protein were mixed with either compound, concentrations ranging from 10^{-10}M to 10^{-3}M . The reaction mixtures were incubated for 2h at RT and free radioligand was removed by filtrating over a Whatman GF/C filter. Filters were dried and 40 μl of scintillation fluid (MicroScintTM-O, Perkin Elmer) was added, radioactivity (cpm) retained on the filters was determined in a Wallac MicroBeta TriLux scintillation counter. The half-maximal inhibitory concentrations (IC₅₀) for these compounds were calculated from normalized dose-response curves obtained using a one site competition binding model (nonlinear regression analyses) of the GraphPad Prism software program. Each assay was performed in triplicate.

For each compound to test, two comparative assays were performed. A first assay on a β_2 -adrenoreceptor-Nanobody fusion locked in its active state by a G protein mimicking Nanobody called $\beta_2\text{AR}$ -Nb80 [224] and a second screen relative to the basal state of the same receptor called $\beta_2\text{AR}$ -Nb69 (Nanobody-enabled activity and efficacy screening platform for GPCR modulating compounds. Pardon E., Betti C. et al., manuscript in preparation). Doing so allowed to classify the efficacy of each hit (i.e. agonist, antagonist or inverse-agonist) [256].

4.4 Results

4.4.1 Creation of the datasets

4.4.1.1 Compatible building blocks

The *frag now* dataset of the ZINC database [253] was processed employing a python script written using the rdkit library [257] in order to retrieve all building blocks compatible with reductive amination and the core fragments. This yielded 18'785 compatible building blocks, i.e. aldehydes or ketones.

4.4.1.2 Surrogates

Every compatible building block was then converted into the corresponding surrogate by the means of a python script written using the rdkit library. Reaction centers (the nitrogen of the introduced amine group) of the surrogates were flagged in order to facilitate filtering later on. The 18'785 compatible reactants were converted into a new set of 26'892 surrogates. The increase in number resulted from the exhaustive enumeration of all possible stereoisomers.

4.4.1.3 Products

The top 500 surrogates were attached to each core fragment according to the reaction scheme for reductive amination. This yielded equally many (500) derivative products.

4.4.2 Fragment docking

Among the 26'892 surrogates, 15'702 were successfully docked to the SBP of the receptor in the active conformation (PDB ID = 4lde) yielding a total of 814'369 docking poses. The same docking procedure was applied for the inactive conformation (PDB ID = 2rh1) and 16'297 surrogates were successfully docked, yielding 1'011'929 poses. These numbers intuitively make sense, since in the inactive conformation, the entrance of the orthosteric site (included in the SBP) is wider than in the active conformation.

4.4.3 Prediction of the binding mode of the core fragments

4.4.3.1 AC molecules

The growing procedure started from docked poses of the four AC fragments identified in the "nanobody screen". We note that no direct interactions with the polar serines at the bottom of the orthosteric site were formed in the predicted poses for these compounds. Three of the core fragments contained an acceptor group, but they were too far ($>4 \text{ \AA}$) away to form direct polar interactions with the serines: the pyrimidine moiety of Z12370550, the thiazole of Z12370253 and the benzothiophene of Z00064947. All the charged amines of the core fragments were found to interact with Asp113^{3,32} and Asn312^{7,39}. The aromatic moieties of the core fragments made hydrophobic interactions with Phe289^{6,51}. As shown in Figure 4.10, those fragments were predicted to bind to the orthosteric site.

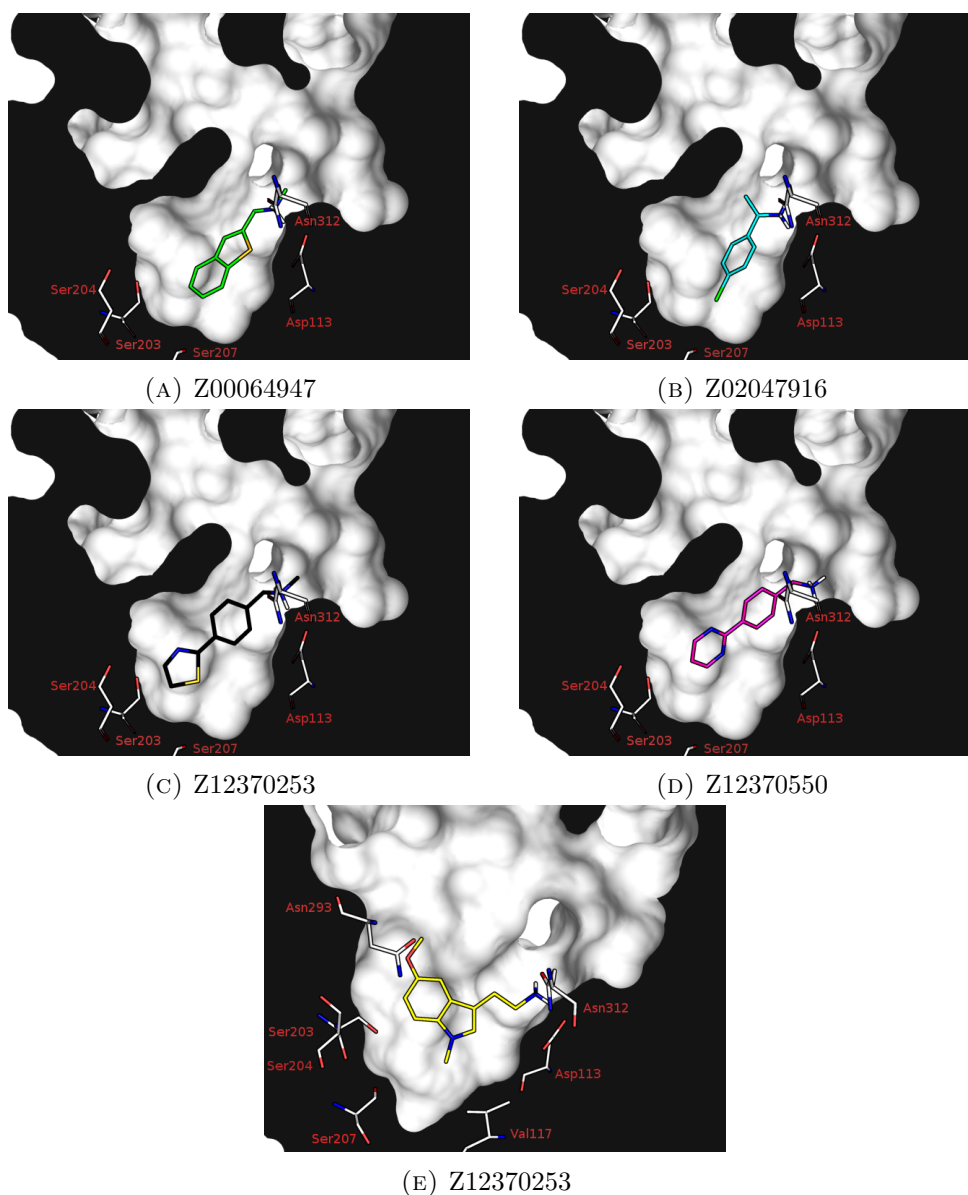


FIGURE 4.10: (a-d) Prediction of the binding mode of the four agonist core fragments within the β_2 AR active conformation. (e) Prediction of the binding mode for Z32501319 after docking to the β_2 AR structure in an inactive conformation.

4.4.3.2 IAC molecules

The second growing procedure started from the docked pose of Z32501319, which is consistent with a pose expected for an IAC binder of the β_2 AR: the charged amine moiety interacts with Asp113^{3,32} and Asn312^{7,39}, the indole moiety makes hydrophobic interactions with Phe289^{6,51}, the methyl on the indole interacts with both Val114^{3,33} and Val117^{3,36}, and the ether moiety engages Asn293^{6,55} in a polar hydrogen bond. As shown in figure 4.10 (e), we predicted Z32501319 to bind to the orthosteric site with room to grow towards the SBP.

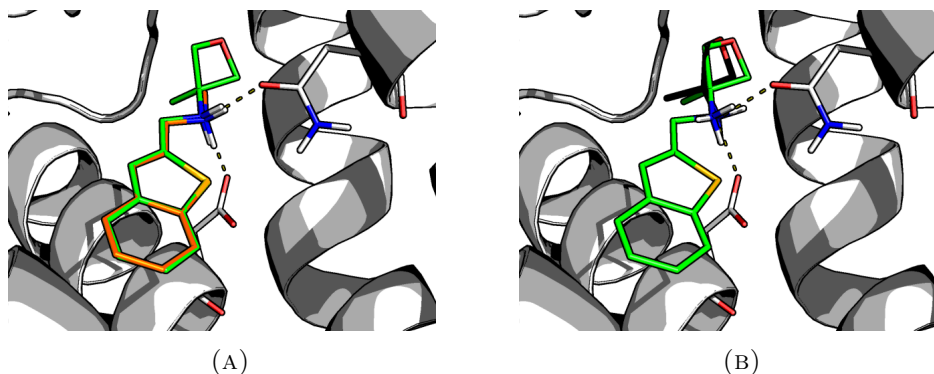


FIGURE 4.11: Example of a product keeping its orientation relative to the poses of the constitutive building blocks. Both the core fragment (orange carbons (A)) and the surrogate (black carbons (B)) are predicted to overlap with the generated product (green carbons).

Product	Core	IC ₅₀ (core)[μ M]	IC ₅₀ (product)[μ M]	Improvement
K011FC001	Z12370253	22	53	0.4
K011FC002	Z12370253	22	96	0.2
K011FC004	Z12370550	79	4.5	17.6
K011FC006	Z02047916	34	7.9	4.3
K011FC008	Z00064947	44	11 000	0.004
K010FC006	Z32501319	20.6	0.53	38.8
K010FC007	Z32501319	20.6	1.05	20
K010FC008	Z32501319	20.6	81	0.25

TABLE 4.1: Summary of the activity of the core fragments and their respective derivative products.

4.4.4 Structure-based screening of the derivative products

The derivative products of every core fragment were docked and visually inspected in order to narrow down the number of candidates for possible further synthesis. The selection of the best derivative products was based on several criteria: docking score, shape complementarity with the receptor, hydrogen bonds to the receptor, overlap of the products with the respective core fragment and surrogate (Figure 4.11) and chemical diversity. Eight ligands not violating these criteria were selected and synthesized (Figure 4.12).

4.4.5 Radio ligand displacement assay

The summary of the radio ligand displacement assay for the eight products is illustrated in table 4.1. The IC₅₀ curves for four products is illustrated in figure 4.13 and the remaining one are in the SI.

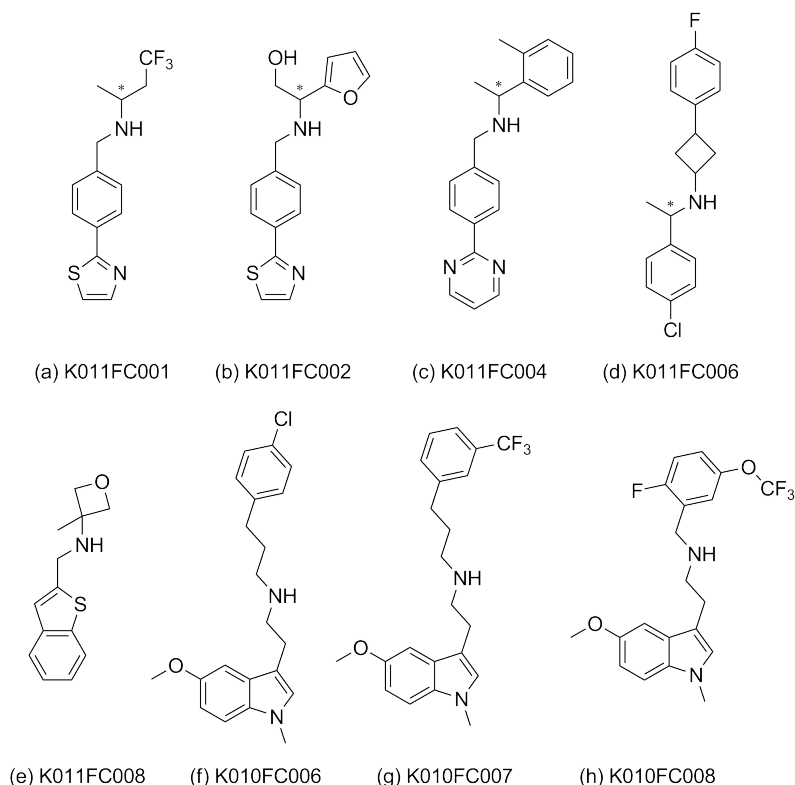


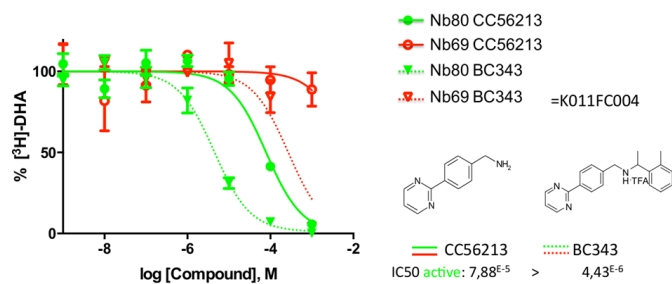
FIGURE 4.12: Eight products were synthesized: five molecules appearing as AC and three as IAC molecules, respectively.

4.5 Discussion

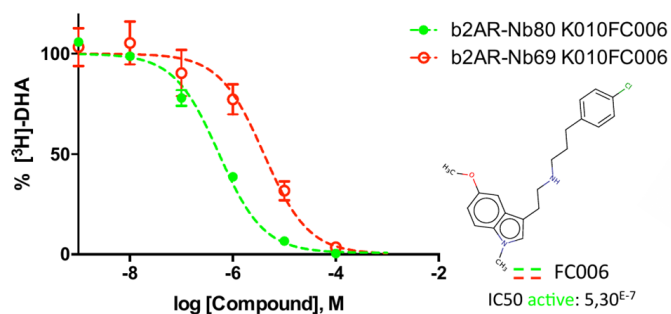
In the following discussion, only the predicted interactions of the synthesized products in the SBP will be detailed, as the interactions in the orthosteric pocket remained constant, as required by our initial hypothesis.

4.5.1 K011FC004

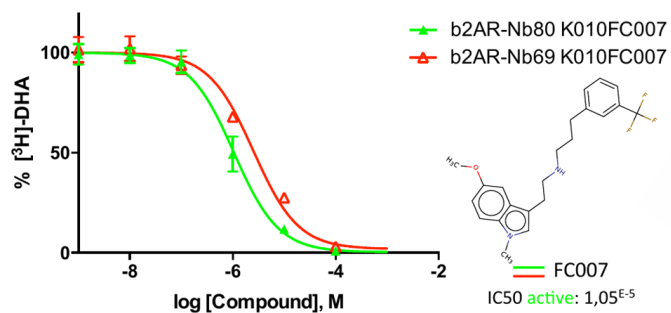
K011FC004 also contains a chiral center and was selected for synthesis because it contains a toluene moiety that can make nice hydrophobic interactions. The R form was predicted to nicely fill the SBP with hydrophobic contact with Trp109^{3,28} and Ile309^{7,36}. On the other hand, the S form is predicted to introduce steric constraints in the SBP forcing the amine out of range from the vital Asp113. Thus, the S form is left with two free H bond donors which is highly unfavorable. K011FC004 was tested as racemic mixture and had an improved affinity (4.5 μ M) compared to its initial core fragment Z12370550 (79 μ M). The docking score predicts the R form to be more favorable than the S form, suggesting that the affinity of the R form could actually be lower than the one measured.



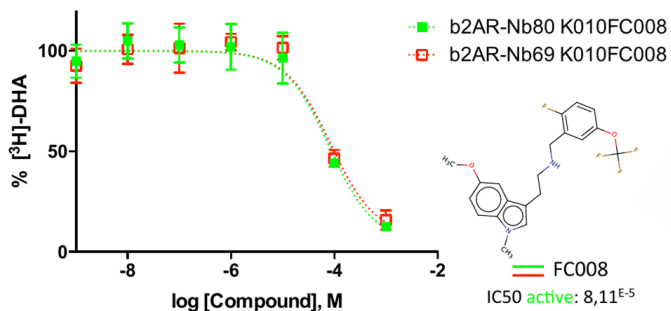
(A) K011FC004



(B) K010FC006



(C) K010FC007



(D) K010FC008

FIGURE 4.13: IC_{50} curves from radio ligand displacement assay. The green curves correspond to the assay made on the active-locked conformation of the receptor (Nb80), while the red curves correspond to the inactive-locked conformation of the receptor (Nb69).

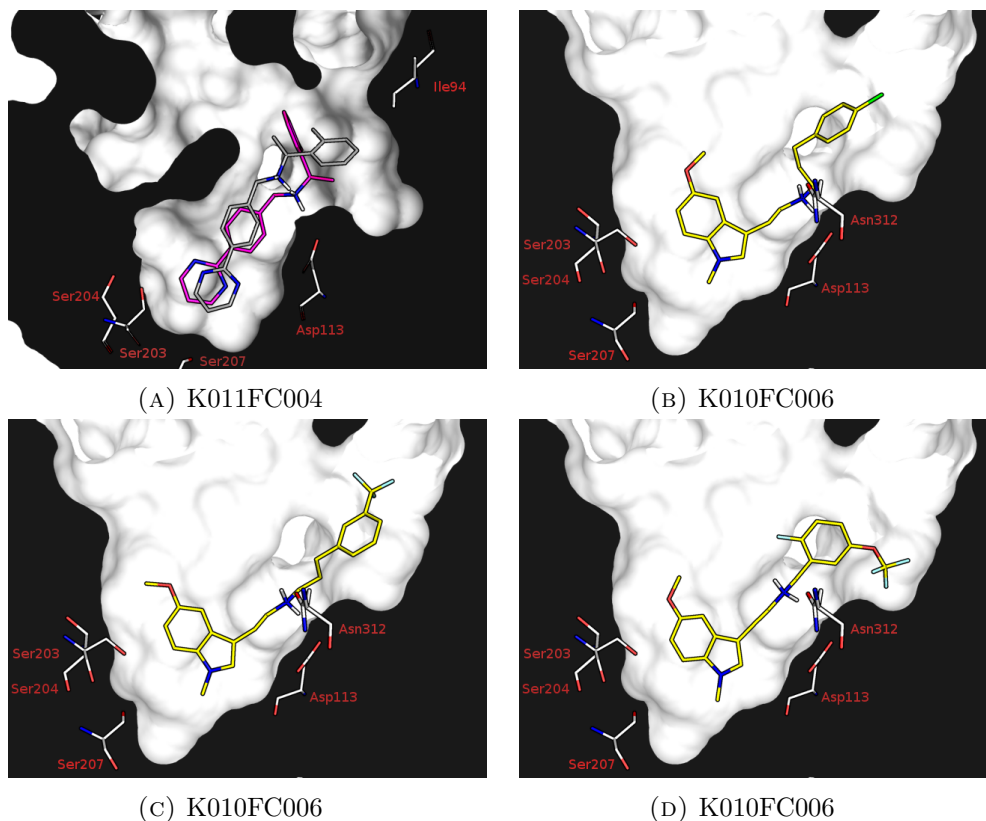


FIGURE 4.14: Predicted binding mode of the remaining products. (a) K011FC004. (b) K010FC006. (c) K010FC007. (d) K010FC008.

4.5.2 K010FC006, K010FC007 and K010FC008

The K010 serie aimed to explore different substitution on the benzene ring when positioned in the SBP, and the importance of the flexible chain between two rigid scaffolds. Two compounds have in common a long flexible chain (6 rotors), while K010FC008 has 4 rotors.

K010FC006 was tested and had a better affinity ($0.53 \mu\text{M}$) compared to its initial core fragment ($20.6 \mu\text{M}$). This important increase in affinity (40 fold) could be explained by strong hydrophobic interaction with Trp109^{3,28} and Ile309^{7,36}.

K010FC007 was tested and had a better affinity ($1.05 \mu\text{M}$) compared to its initial core fragment ($20.6 \mu\text{M}$). K010FC007 contains a trifluoro moiety which is more bulky than K010FC006. The docking pose suggest that K010FC007 could also make hydrophobic contacts with Trp109^{3,28} and Ile309^{7,36}.

K010FC008 was tested and had an decreased affinity ($81 \mu\text{M}$) compared to its initial core fragment ($20.6 \mu\text{M}$). This small decrease in affinity (4 fold) can be explained by the the size of the extension which introduce more steric constraints in the SBP which probably disrupt the hydrophobic interaction with Trp109^{3,28}. K010FC008 has two less

rotors than its two aforementioned analogs, and this could suggest that K010FC008 has thus less degree of freedom in order to accommodate into the SBP and its interactions are less optimized.

4.6 Conclusions

Starting from diverse fragments exhibiting activity against the β_2 AR, we were able to grow those fragments guided by the protein structure. For several ligands we were able to improve the initial affinities. The *in silico* workflow we developed for this purpose tackles three challenges at the same time. First, the fragments need to be grown towards a compatible region offering enough space for the addition of another building block. Secondly, the suggested molecules ought to be very likely synthesizable. Lastly, the molecules should also have a high probability of being active against the target. PINGUI does this by focusing on few candidates and pruning search space at an early space.

All the initial fragments in this study had two points in common: they were predicted to bind in the orthosteric site and they all contain an amine interacting with Asp113^{3.32}. Thus, an extension of the fragments towards the SBP was feasible. In the interest of focusing on molecules that are likely to be synthesizable, we used a library of 58 robust organic reactions [1] and applied them to our fragments. We decided to focus on products obtained through reductive amination, because the amine was the reaction center and all the generated products could potentially extend past Asp113^{3.32} to fill the SBP. All eight molecules that we suggested were indeed successfully synthesized, highlighting that reductive amination can be regarded as a robust reaction.

In the interest of increasing the chance of our suggestions being active, our workflow prioritized extensions that already entertain favorable interactions with the protein. A two-step docking of first the surrogate and then the resulting product allowed to increase the robustness of our prediction. If a promising product deviated too much from its initially predicted position, it was not pursued further. We are aware that there are retrosynthetic experimental analyses, including X-ray crystallography of the resulting fragments, that show that such binding mode faithfulness is not always the case [258]. However, one can afford to be more elitist in this approach, as combinatorics works in our favor by generating way too many solutions that need to be pruned. Moreover, docking is notorious for producing many false positives [259, 260], and applying stringent selection criteria oftentimes improves true positive rates. In our hands, these considerations allowed us to significantly improve half of the original fragment binders. It is difficult to assess whether this is better or worse than what a human medicinal chemist might

have achieved. However, when one takes a look at current literature examples, an improvement rate of 50 % seems higher than what is usually reported.

The presented workflow is not strictly automatized, because it relies on spatial reasoning abilities during the evaluation of the poses. However, since most of the growing steps rely on computational approaches, we decided to make them freely available to the public. Thus, we created PINGUI, a toolbox that we hope will help medicinal chemists as an idea generator in fragment-based ligand discovery project. PINGUI is available online and is implemented alongside the SCUBIDOO database [261]. PINGUI features many options that were instrumental in this project, namely displaying the growing vectors for any given fragment, the creation of customized derivative libraries or deconstructing a molecule into smaller fragments based on known reactions. The last feature allows to fully utilize a construction/deconstruction approach that can be very helpful in an FBDD project.

As with SCUBIDOO, we do not see the utility of this approach so much in the probability that it will yield a highly potent compound, but rather in providing medicinal chemists with creative suggestions. Hence, we believe that PINGUI is best employed at the very early stages, when SAR needs to be generated quickly and with limited synthetic efforts. We also think that it will enable researchers who are not dyed-in-the-wool organic chemists (such as ourselves) to develop and synthesize comparatively simple substances, thereby speeding up research and potentiating the availability of pharmacologically active chemical matter.

4.7 Supporting Information

4.7.1 Radio ligand displacement assay

The IC₅₀ curves from the radio ligand displacement assay of compounds K011FC001, K011FC002, K011FC006 and K011FC008 are gathered in picture 4.15

4.7.2 Discussion

4.7.2.1 K011FC001

K011FC001 exists in two enantiomers. The *S* form was predicted to make mostly apolar interactions in the SBP, with the trifluoro moiety interacting with Ile309^{7.36}, Ile94^{2.65} and Trp109^{3.28} (Figure 4.16). The charged amine makes polar interactions with Asp113^{3.32} and Asn312^{7.39}. Interestingly, the *R* form keeps the overall binding mode except for the amine which flips, thus breaking the polar bond with Asp113^{3.32}. Thus, the *R* form presents an unsatisfied hydrogen bond donor. K011FC001 was tested as racemic mixture and had a lower affinity (53 μ M) compared to its initial core fragment Z12370253 (22 μ M). The docking predictions scored the *S* form more favorable than the *R* form, hence the affinity of one enantiomer could potentially be lower than the apparent IC₅₀ of 53 μ M. However, because of the low affinity, we decided not to attempt purification of the individual enantiomers.

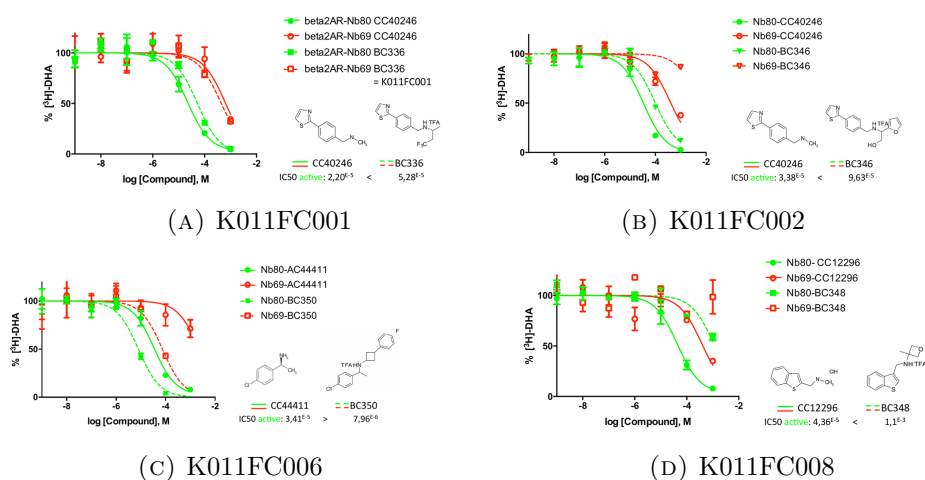


FIGURE 4.15: IC₅₀ curves from radio ligand displacement assay. The green curves correspond to the assay made on the active-locked conformation of the receptor (Nb80), while the red curves correspond to the inactive-locked conformation of the receptor (Nb69).

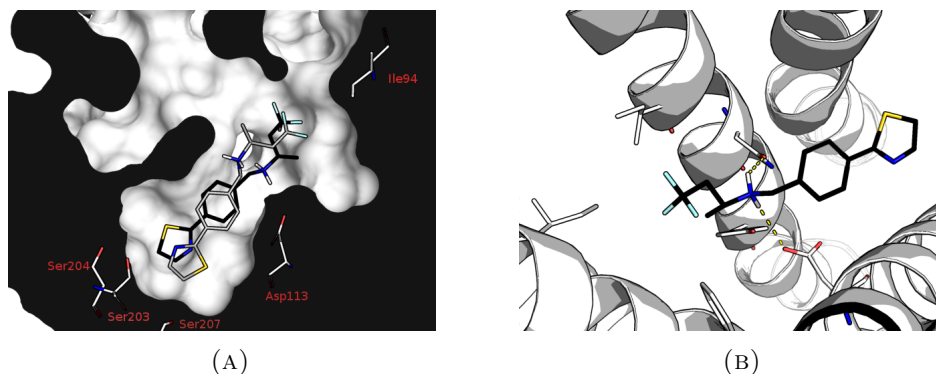


FIGURE 4.16: Prediction of the binding mode of K011FC001 using docking. (a) Sliced view of the binding site. The *S* form (black carbons) interacts with Asp113^{3.32} and Asn312^{7.39} while the *R* form (grey carbons) can not. (b) Side view from the SBP highlighting the hydrophobic interactions of the trifluoro moiety of the *S* form with Ile309^{7.36}, Ile94^{2.65} and Trp109^{3.28}. Polar interactions of the charged amine with Asp113^{3.32} and Asn312^{7.39} are represented in yellow dashed lines.

4.7.2.2 K011FC002

K011FC002 contains a chiral center too, and was selected for synthesis because it contains a hydroxy moiety that was predicted to make a hydrogen bond with Tyr316^{7.43} (figure 4.17), albeit from the opposite side of the nitrogen compared to adrenaline. This motif was quite frequent among all the suggested products, and since no publicly known active molecules contain it, we wanted to explore its pharmacological relevance. The hydroxy moiety of the *R* form was predicted to form a hydrogen bond with Tyr316^{7.43}, while the furan makes hydrophobic contacts with Ile309^{7.36}. On the other hand, the *S* form is predicted to lose all key polar interactions: the charged amine as well as the hydroxy group are too far from Tyr316^{7.43} and Asp113^{3.32}. Thus, the *S* enantiomer is left with two free hydrogen bond donors which is highly unfavorable. K011FC002 was tested as racemic mixture and had a lower affinity (96 μM) compared to its initial core fragment Z12370253 (22 μM). As before, the docking score predicts the *R* form to bind with a more favorable affinity than the *S* form, suggesting that the affinity of the *R* form could actually be lower than the one measured.

4.7.2.3 K011FC006

K011FC006 was selected because the initial core fragment was predicted to flip its binding mode therefore interacting in the SBP with Trp109^{3.28} and Ile309^{7.36}. K011FC006 contains only hydrophobic moieties except for the charged amine. K011FC004 was tested and had a better affinity (7.9 μM) compared to its initial core fragment Z12370550 (34 μM). This slight improvement for this product could be explained by a better desolvation due to its hydrophobic nature.

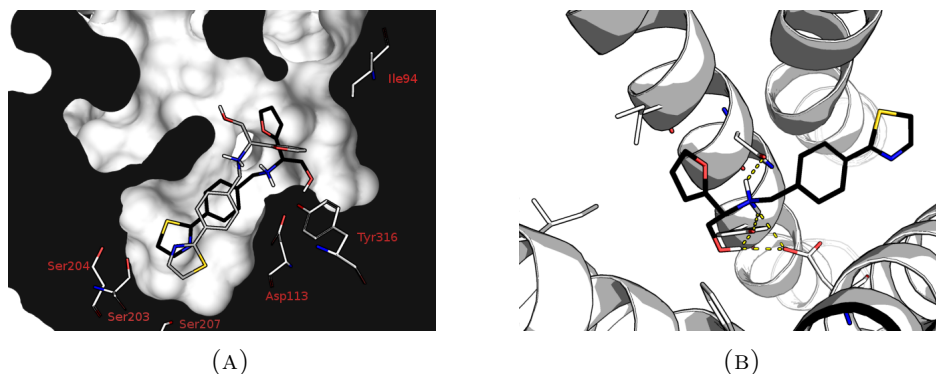


FIGURE 4.17: Prediction of the binding mode of K011FC002 using docking. (a) Sliced view of the binding site. The *S* form (black carbons) makes polar interactions with Asp113^{3.32}, Asn312^{7.39} and Tyr316^{7.43} while the *R* form (grey carbons) can not. (b) Side view from the SBP highlighting the hydrophobic interactions of the furan moiety of the *S* form with Ile309^{7.36}. Polar interactions of the charged amine with Asp113^{3.32}, Asn312^{7.39} and Tyr316^{7.43} are represented in yellow dashed lines.

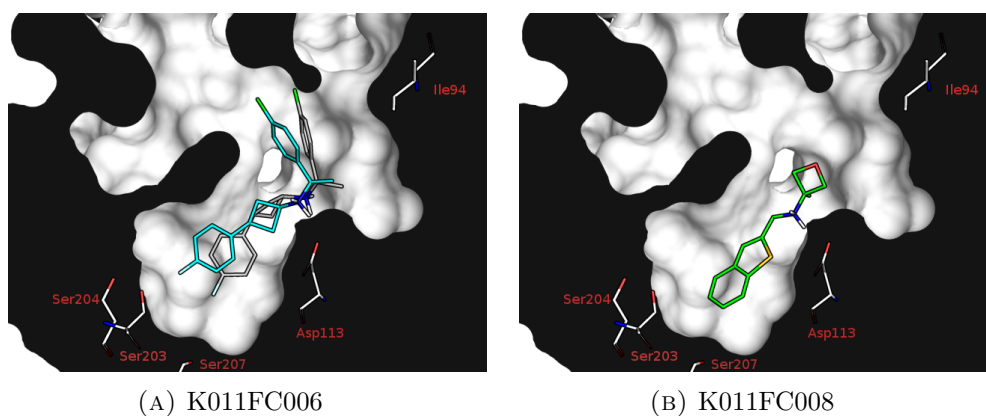


FIGURE 4.18: Predicted binding mode of the remaining products. (a) K011FC006. (b) K011FC008.

4.7.2.4 K011FC008

K011FC008 was selected due to the small size of its extension (only four new heteroatoms) in comparison to the other products and also because it did not contain an aromatic moiety. K011FC008 showed an important fall in affinity (250 fold) which could partially be explained by the lack of strong hydrophobic contacts with Trp109^{3.28} or Ile309^{7.36}. Furthermore the ether moiety introduces two free lone pairs that will induce a small desolvation penalty.

"There is a beast in every man and it stirs when you put a sword in his hand."

Ser Jorah Mormont, *Game of Thrones*, season 3 episode 3 (2013)

Chapter 5

SCUBIDOO

5.1 Article

SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability

F. Chevillard and P. Kolb*

Department of Pharmaceutical Chemistry, Philipps-University Marburg, 35032 Marburg, Germany

S Supporting Information

ABSTRACT: De novo drug design is widely assisted by computational approaches that enable the generation of a tremendous amount of new virtual molecules within a short time frame. While the novelty of the computationally generated compounds can easily be assessed, such approaches often neglect the synthetic feasibility of the molecules, thus creating a potential hurdle that can be a barrier to further investigation. Therefore, we have developed SCUBIDOO, a freely accessible database concept that currently holds 21 million virtual products originating from a small library of building blocks and a collection of robust organic reactions. This large data set was reduced to three representative and computationally tractable samples denoted as S, M, and L, containing 9994, 99 977, and 999 794 products, respectively. These small sets are useful as starting points for ligand identification and optimization projects. The generated products come with synthesis instructions and alerts of possible side reactions, and we show that they exhibit drug-like properties while still extending into unexplored quadrants of chemical space, thus suggesting novelty. We show multiple examples that demonstrate how SCUBIDOO can facilitate the search around initial hits. This database might be a useful idea generator for early ligand discovery projects since it allows a focus on those molecules that are likely to be synthetically feasible and can therefore be studied further. Together with its modular building block construction principle, this database is also suitable for structure–activity relationship studies or fragment-growing strategies.



INTRODUCTION

Chemical space is vast. The question is how to navigate it in order to identify ligands that can serve as modulators for pharmaceutically interesting targets. It seems likely that chemically not-yet-realized molecule sets hold many potent ligands for a variety of targets. At present, the *in silico* realm is the only place where we can hope to enumerate molecules that might be stable under ambient conditions. Such efforts have been undertaken, pioneered by Lederberg.¹ Currently, the most advanced development comes from the Reymond lab with their chemical universe database GDB,² which in its current incarnation enumerates virtual molecules containing up to 17 heavy atoms.

However, for all such virtual databases, the critical point is the actual synthesis of the generated molecules. Despite following strict chemistry rules, such molecules might turn out to be unsynthesizable with reasonable effort. This becomes a barrier in the initial stages of a lead-finding project, where a quick go/no-go decision is desired. In addition, automation of synthesis protocols is a currently intensively investigated topic of research,^{3–5} and new open-innovation initiatives have arisen aiming at discovering novel chemical entities.⁶ Thus, providing suggestions for further synthetic developments could improve such protocols even more.

Another problem of such enumerated databases is their sheer size: GDB currently holds 166 billion molecules,⁷ which basically is computationally intractable except for ultrafast two-dimensional methods. For structure-based methods, such as docking, this is unfeasible at the moment.⁸

In order to advance on both topics, we have created the Screenable Chemical Universe Based on Intuitive Data OrganizatiOn (SCUBIDOO) and made it freely available to the general public. The current version was obtained by exhaustively reacting a set of building blocks with 58 highly reliable reactions. Such an approach is not completely novel^{9–15} but has rarely been carried out entirely outside of an industrial framework.¹⁶ The set of 58 reactions is the work of Hartenfeller et al.¹⁷ and represents the most commonly used reactions in the pharmaceutical field. The authors compared their collection with a study by Roughley and Jordan¹⁸ and showed that the 58 reactions cover 48.3% of the 7315 reaction steps described in this review. In a later publication, the authors investigated the coverage of chemical space afforded by their 58 robust reactions when applied to 26 043 common molecular building blocks.¹² They generated a limited number of one-step synthesis products by combining every building block with a maximum

Received: April 13, 2015

of 20 reaction partners for each reaction. This protocol yielded a data set of 1 696 226 closed-source products and revealed that they were able to successfully reconstruct known ligands and sample the chemical space of bioactive compounds in a wide range of target families. Furthermore, they suggested that there is still a vast amount of unexplored bioactive space, which could contain “low-hanging fruit”.

Going beyond this study, with SCUBIDOO we have now started to exhaustively react building blocks with each other in order to completely cover the chemical space thus accessible. This exhaustiveness comes at a price, however: even when starting with a relatively modest number of building blocks (~8000), this exhaustive reaction scheme yields a large number of lead-like molecules (more than 21 million) in the end. This is desired in the sense that the larger this number is, the greater is the amount of chemical space we can cover. However, many millions of virtual compounds make the utility for computational approaches dubious again. To address these divergent tendencies, we make use of stratified sampling to provide a representative subset of the database. Consequently, a user of our freely accessible database can obtain primary ligand candidates from a small and processable sample through virtual screening, pick those that can be synthesized with relative ease, and advance from hits efficiently by searching all analogues in SCUBIDOO. Since every molecule in the database comes with synthesis instructions, information about potential side reactions, and alternative synthetic pathways, it represents a fast and efficient way to start on a new target. SCUBIDOO can thus help to probe this unexplored potentially bioactive space more intuitively.

Moreover, we think that SCUBIDOO will help to fight “molecular obesity”,¹⁹ defined as the steady increase in the molecular size of drug candidates during medicinal chemistry development, since it facilitates starting a ligand discovery project focused on fragment considerations. Of course, the database as such can never be complete, but the concept is amenable to expansion.

In this article, we describe the development of the SCUBIDOO concept and a first database of 21 million compounds. We then show that the obtained database contains products comparable to currently existing drugs as well as databases of lead-like molecules of similar size. Still, SCUBIDOO extends into different quadrants of chemical space, as we demonstrate through principal component analysis (PCA). Moreover, we show the usefulness of SCUBIDOO through several examples in which we embark from close analogues of drug candidates or ligands and harvest even closer analogues or existing active compounds within a few mouse clicks. We also show that in those cases where synthetic information was publicly available, the reactions used to obtain these molecules match the ones suggested in SCUBIDOO.

METHODS

Reactions. The list of 58 reactions is provided in SMARTS notation in the study of Hartenfeller et al.¹⁷ and in Table S1 in the [Supporting Information](#).

Data Sets. Reactants: Building Blocks. Any product in SCUBIDOO is generated by combining a maximum of two building blocks. The initial set of 18 561 building blocks was downloaded from the ChemBridge Web site.²⁰ By means of a Python script written using the RDKit library,²¹ routine filters were then applied to the building blocks to strip counterions and remove duplicates. In order to avoid overly complex

reaction products that might necessitate more complex synthesis strategies, the reactant library was also pruned to narrow the range of generated products. The following criteria were used in the filters:

- $MW \leq 250$ Da. Products will thus mostly be below 500 Da.
- Number of *rotatable bonds* ≤ 2 . Since a reaction can introduce one or even two new rotatable bonds, this filter restrains the products to a low number of rotors (with a maximum of 6). Doing so makes the use of structure-based strategies such as docking more reliable, since the estimation of the binding energy of molecules with a high number of rotors is prone to fail.²²
- Number of *chiral centers* ≤ 1 . This filter was introduced with one goal: to facilitate synthesis. Since some of the 58 reactions introduce a chiral center into the resulting product, this limit on the number of chiral centers yields products with a maximum of three chiral centers.

Reactants: Analysis of Reagent Classes. In a recent study, Goldberg et al.²³ highlighted the importance of building block libraries in the interest of improving compound quality. They also introduced a classification of building blocks into 23 reagent classes based on functional groups. We used this class attribution in the analysis of the composition of the ChemBridge data set in this study, employing a Python script written using the RDKit library. Every building block was assigned to at least one reagent class but could belong to several classes. Using the SMARTS-encoded reagent class definitions as provided by the authors,²³ we reduced the list of 23 reagent classes to 22 by merging the *benzaldehyde* and *heterocyclic benzaldehyde* classes into a single *aromatic aldehyde* class. The list of 22 reagent classes is available in Table S2 in the [Supporting Information](#). We note that this classification was done only to investigate the diversity of the library and had no influence on the reactions that each fragment was able to undergo.

Products: Screenable Chemical Universe. The filtered building blocks were then exhaustively reacted against each other via the 58 reactions using an in-house Python script written using the RDKit library. This procedure can be divided into three loops: the first one over all of the building blocks (B1), the second one over all of the reactions, and the third one over all of the building blocks (B2) ([Figure 1](#)). All of the products were charge-neutralized in order to simplify subsequent steps. Duplicate products were filtered according to their isomeric canonical SMILES notation and the reaction involved in the synthesis. We deliberately wanted to keep track

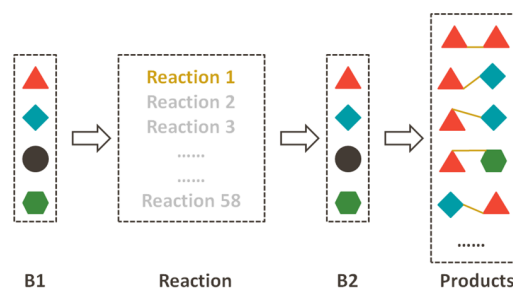


Figure 1. Schematic depiction of the creation of the SCUBIDOO database. B1 and B2 are building blocks which are then connected by all compatible reactions.

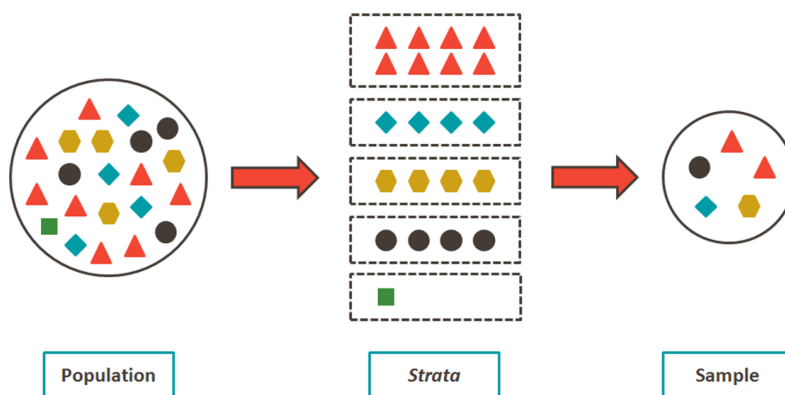


Figure 2. Stratified balanced sampling algorithm.

of multiple synthesis routes for the same product in order to display them on the subpage of each product as alternative synthetic routes. This is valuable information for chemists and increases the chances of synthesizing a particular compound. Afterward, the products were filtered using the PAINS filter level A²⁴ in order to get rid of compounds that have a high chance of generating artificial results during follow-up biological assays. Stereoisomers were generated using *flipper*²⁵ in order to enumerate all possible products in cases where a reaction introduced a new chiral center.

Representative Samples: Stratified Balanced Sampling. Since the present version and therefore also all future larger versions of SCUBIDOO are too big to be rapidly processed with structure-based approaches, we reduced it to three representative samples of different sizes (S, M, and L). This provides users with optimal sets for different applications. This procedure was done using the *cubestratified*²⁶ algorithm of the *balancedSampling* package²⁷ within the R statistics environment.²⁸ The stratified balanced sampling approach is a popular algorithm used for population surveys, allowing extraction of a representative sample. The algorithm consists of two stages (Figure 2). In the first stage, *stratification*, the studied set is divided into subgroups called *strata*. In this study, the strata are defined by the reactions, and each product is assigned to exactly one stratum. In a second step, *balanced sampling*²⁹ is applied within each stratum in order to select representative products. This selection is based on auxiliary variables defined as chemical descriptors (here, molecular weight, logP, number of H-bond donors, number of H-bond acceptors, and topological polar surface area were used) and aims to reflect the overall composition of each stratum. Furthermore, the sample size of each stratum is proportional to its total size. It is important to mention that balanced sampling does not guarantee that all of the strata defined initially are present in the final sample in cases where the strata sizes vastly differ. This algorithm is exceptionally fast even for huge amounts of data³⁰ and is thus well-suited for the processing of our screenable chemical universe (and its future larger incarnations).

DrugBank. DrugBank version 4.1 was downloaded from its Web site³¹ for comparison purposes. Only approved and experimental drugs with molecular weights lower than 500 Da were kept. This led to a data set of 1510 molecules.

ZINC: Lead-like Subset. The lead-like data set was downloaded from ZINC.³² Since this data set is quite large (more than 6 million compounds), a sample of 10 000 compounds was randomly selected using the *sample* function within the R statistics framework.

PDB: Ligands. All of the ligands present in the Protein Data Bank (PDB) were downloaded from its Web site.³³ The ligands were then filtered according to molecular weight (≤ 500 Da) and the number of rotatable bonds (≤ 6) in the interest of narrowing down the ligands to molecules close to the products of SCUBIDOO. This yielded 17 140 ligands.

Analogues of DB08235. The five analogues of DB08235 were prepared for docking using OMEGA,³⁴ with a maximum of 1000 conformers. The protonation states were defined using QUACPAC.³⁵

Ligand-Based Application: Similarity Screening. A ligand-based screening strategy was applied in order to retrospectively assess the usefulness of SCUBIDOO. Two data sets were used in this comparison: DrugBank and all of the ligands extracted from the PDB. Each of these data sets was compared to the three samples of SCUBIDOO using FCFP4 fingerprints.³⁶ For each product in each sample, the most similar drug or ligand according to the Tanimoto coefficient was retrieved. All pairs with a Tanimoto score higher than 0.6 were visually inspected in order to identify the representative examples described in this article.

Synthetic Accessibility. To assess the synthetic feasibility of the products within SCUBIDOO with an alternative method, the synthetic accessibility (SA) score³⁷ was computed for each of the products using an RDkit-based Python script. SA score estimation is based on fragment contributions and a complexity penalty (chiral centers, weight, large rings). SA scores range between 1 and 10, with 1 indicating a simple molecule that should be easy to make and 10 representing a complex molecule that is likely to be hard to synthesize.

Principal Component Analysis (PCA). DrugBank, the lead-like subset of ZINC, and the SCUBIDOO S sample were compared using the PCA function as implemented in the R statistics environment. The descriptors used were molecular weight, logP, number of H-bond donors, number of H-bond acceptors, number of rotatable bonds, topological polar surface area, and the Bertz index,³⁸ which estimates the molecular complexity.

Web Interface. The database is freely accessible at www.kolblab.org/scubidoo. All of the chemical descriptors shown for each molecular entity are computed using the RDKit library. The partition coefficient, logP, is predicted using Crippen's approach.³⁹ TPSA represents the topological polar surface area of the molecule.⁴⁰ Two different searches are available: by product or by building block. These are described in more detail below:

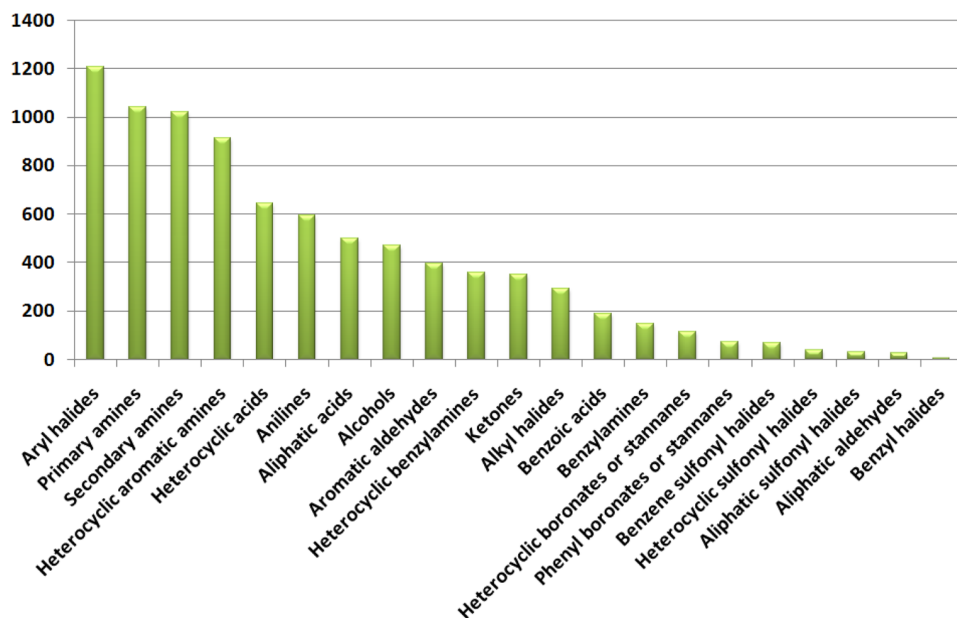


Figure 3. Frequencies of reagent classes in the ChemBridge building blocks data set.

Product Search. One can search for a product in SCUBIDOO using either its ID or its SMILES string. A comprehensive array of information is displayed when the product is retrieved: molecular descriptors, synthetic route, building blocks involved in the formation of the product, possible side reactions, and alternative synthetic routes. An alternative synthetic route is defined as a possibility to obtain a given product using a different reaction or a different pair of building blocks. A simple color classification has been implemented in order to quickly make the user aware of potential problems. *Red* products are ones that still contain reactive features (i.e., an electrophile and a nucleophile) and thus have the potential to react further. *Orange* products are ones for which the building blocks used in the synthesis are involved in more than one reaction (i.e., side reactions might occur). *Green* products are compounds that do not fall into the two aforementioned categories. Reactive features are retrieved using a Python script written using the RDKit library. Nucleophiles are defined as amines, alcohols, and thiols, while electrophiles were defined as acids, halides, and carbonyls. All of the functional groups are encoded as SMARTS and are provided in Table S3 in the [Supporting Information](#).

Building Block Search. A building block search also displays a multitude of information: molecular descriptors, the top four analogue building blocks along with their Tanimoto scores using MACCS fingerprints, and all of the products in SCUBIDOO based on this building block grouped by reaction. For the latter option, the user has the possibility to download the selected products in SMILES format.

RESULTS

Creation of the Data Sets. Filtering of the Building Block Library. The initial library contained 18 354 building blocks. In a first filtering stage, all of the counterions were stripped and duplicates were removed, yielding 14 831 building blocks. Then only building blocks with molecular weights lower than 250 Da were kept, leaving 13 678 entities. Removing building blocks with three or more rotatable bonds reduced the library to 8006

entities. Applying a last filter allowing zero or one chiral centers resulted in a final building block library of 7805 molecules.

Creation of the Product Database. The 7805 building blocks were reacted against each other, generating 17 538 385 products. The computational part was carried out on a cluster of 192 CPUs and took less than 12 h. Duplicate products were removed, yielding 14 215 760 products. Then the PAINS filter level A was applied, leading to a reduction to 14 072 131 products. Afterward, stereoisomer generation was carried out, giving rise to a final database of 21 035 460 products.

Creation of Representative Samples. The 21 million products were reduced to three representative samples of different sizes (S, M, and L) using the stratified balanced sampling algorithm. All of the products were regrouped into strata by reactions, leading to 45 strata, i.e., 13 reactions never occurred. The never-occurring reactions are based on building blocks that are not present in the currently used library. The list of reactions that are not present in SCUBIDOO is provided in Table S4 in the [Supporting Information](#). Next, during balanced sampling, the representative products for each stratum were selected using chemical descriptors (molecular weight, logP, number of H-bond donors, number of H-bond acceptors, and topological polar surface area) as auxiliary variables. In the end, the S, M, and L samples contained 9994, 99 977 and 999 794 compounds, respectively. For analysis, only the S sample was used. Only the L sample contained at least one representative of each reaction; the S sample was missing four reactions and the M sample two reactions.

Analysis of the Data Sets. Analysis of the ChemBridge Building Block Library. The breakdown of the ChemBridge building block library by reagent class is displayed in [Figure 3](#). The most frequent reagent classes, which are primary, secondary, and heterocyclic aromatic amines, heterocyclic acids, and aryl halides, are consistent with the most frequent reactions used during the creation of SCUBIDOO. Indeed, the Schotten–Baumann amide, Buchwald–Hartwig, reductive amination, and Negishi reactions, which are the top four reactions employed in SCUBIDOO, require the aforementioned reagents.

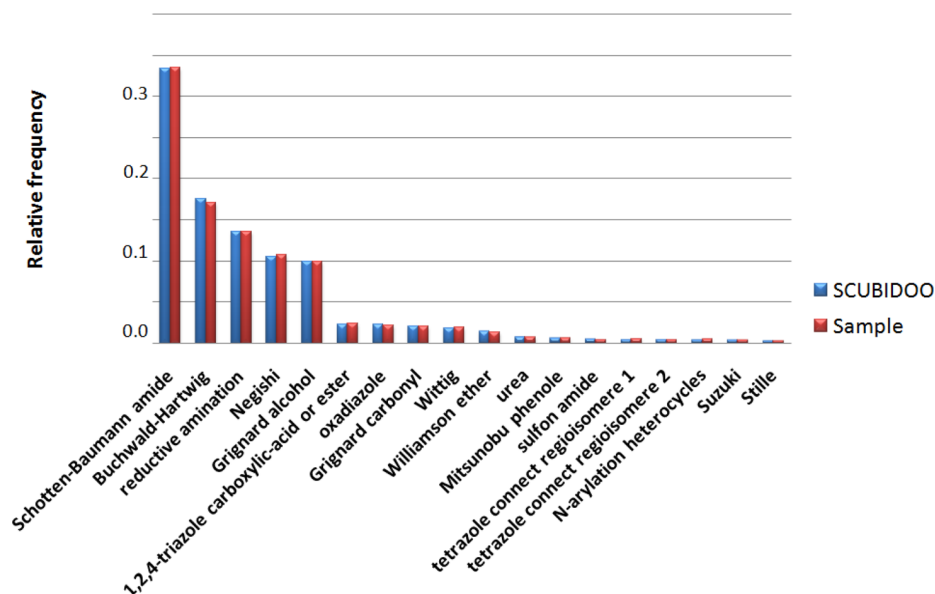


Figure 4. Relative frequencies of the reactions used in SCUBIDOO (blue) and the S sample (red). Only reactions employed at least 50 000 times are represented here.

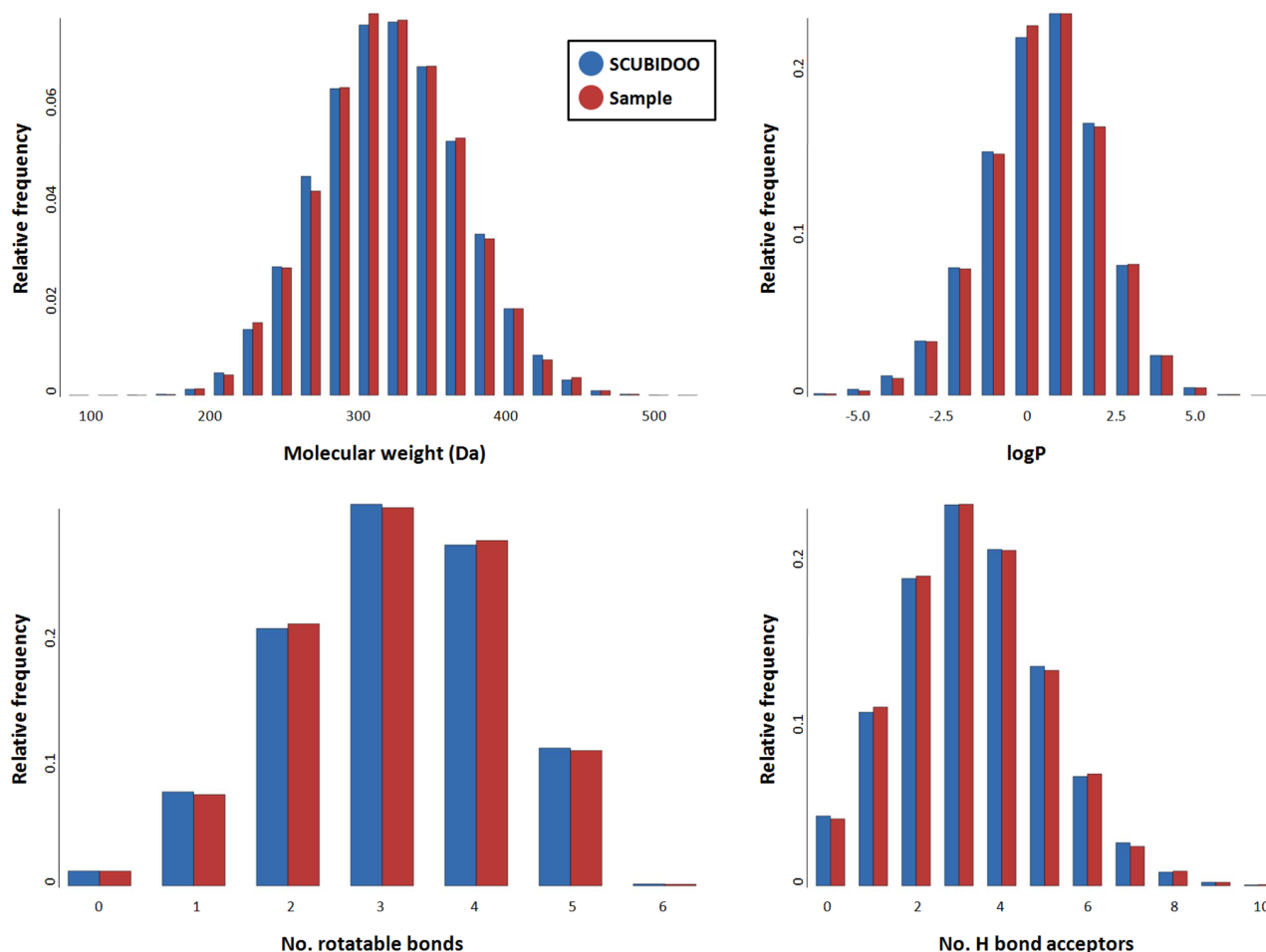


Figure 5. Relative frequencies of the descriptors molecular weight (upper left), logP (upper right), number of rotatable bonds (lower left), and number of H-bond acceptors (lower right) in SCUBIDOO (blue) and the S sample (red).

Frequencies of the Reactions. During the first stage of stratified balanced sampling, all of the products from

SCUBIDOO were regrouped by reaction in order to define the strata. To ensure that the frequencies of the reactions

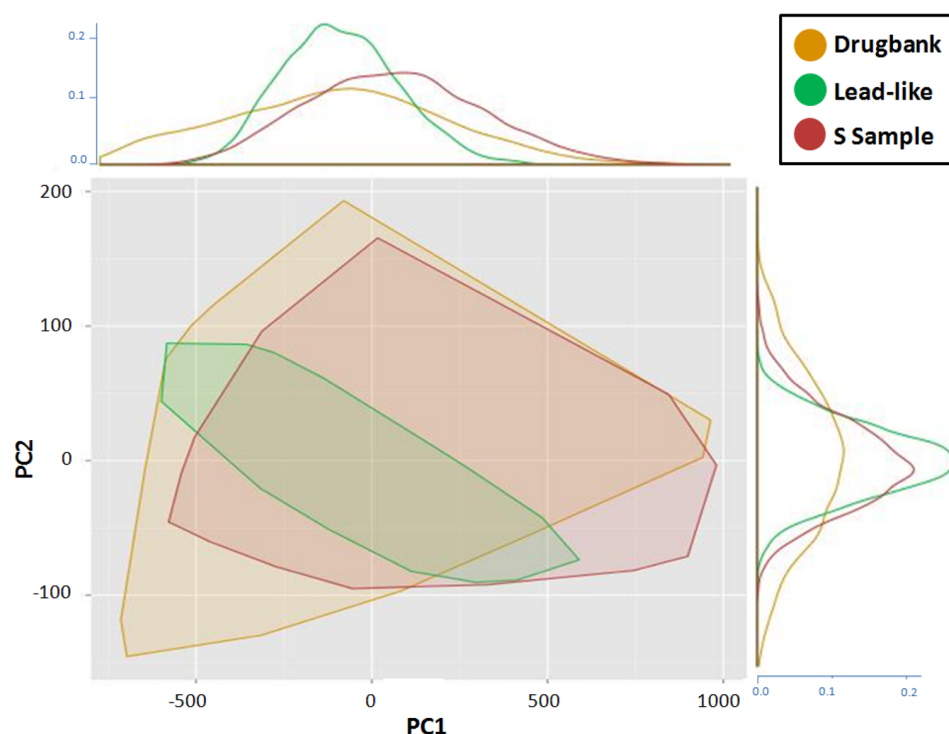


Figure 6. Principal component analysis of DrugBank (orange-yellow), the lead-like subset (green), and the S sample (red). The first principal component explains 34.9% of the total variance, and the second explains 28.3%. The two principal components thus cover 63.2% of the total variance. The marginal histogram on each axis represents the density distribution of each data set according to the same color code.

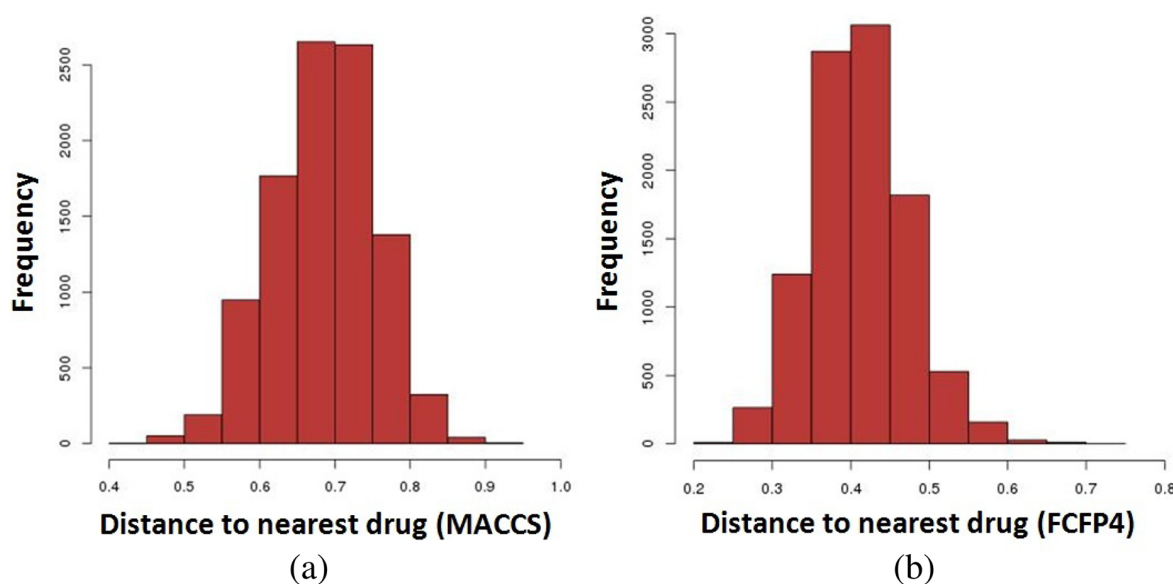


Figure 7. (a) Tanimoto score distributions of the 9994 products in the S sample compared with the 1540 compounds of DrugBank using (a) MACCS keys and (b) FCFP4 fingerprints.

within the entire population and within the S sample were of similar distribution, we plotted the reaction frequencies as a histogram (Figure 4). It is intriguing that almost 75% of the generated products of SCUBIDOO are based on four chemical reactions: Schotten–Baumann amide (33.3%), Buchwald–Hartwig (17.5%), reductive amination (13.6%), and Negishi (10.5%). This partitioning is of course related to the distribution of reagent classes observed in the previous section. It is also similar to the findings of Hartenfeller et al.¹² and lines

up with the study of Roughley and Jordan,¹⁸ where these four reactions are among the top six reactions used in the pharmaceutical field. In contrast, the popular Suzuki coupling is underrepresented here. This is due to the fact that only a few boronic acids (97) were present in the initial building block library.

Chemical Properties. In the second stage of stratified balanced sampling, the representative products of each stratum were selected using auxiliary variables (i.e., molecular

descriptors). For the purpose of comparing SCUBIDOO to the S sample, the distributions of these auxiliary variables for each data set were plotted as histograms (Figure 5). The distributions of the two data sets are similar, indicating that the representative sample respects the heterogeneity of the initial population.

Diversity and Novelty. Analysis of the space spanned by the aforementioned physicochemical properties (Figure 6) using the two main principal components depicts the S sample as overlapping with the property regions of known drugs and lead-like compounds. This suggests that many of the generated products are in principle drug-like. However, despite this overlap, there are also many molecules from the S sample that are located within regions where known drugs and lead-like compounds are absent, indicating the existence of potentially chemically novel compounds. The composition of features of each plane is provided in Table S5 in the Supporting Information.

We also compared the S sample against DrugBank in more detail. For each product from the S sample, the distance to the nearest drug was calculated using MACCS keys⁴¹ and FCFP4 fingerprints employing the Tanimoto coefficient. We used two different fingerprints for this comparison because we wanted to obtain two different opinions on similarity. The MACCS fingerprint contains 166 bits and is used for substructure searching. Each bit position specifically encodes a common functional group. In contrast, FCFP4 fingerprints are topological circular fingerprints, which are not predefined and can represent a large number of different molecular features. Therefore, these fingerprints highlight how particular features of known drugs are retrieved within SCUBIDOO products. Tanimoto score frequencies are plotted in Figure 7. Interestingly, the distribution for MACCS keys similarity is centered around a Tanimoto score of 0.7, suggesting that the S sample contains both similar and dissimilar products in comparison with known drugs. We treat 0.7 as a cutoff between similar and dissimilar compounds for MACCS keys, as it was shown to have discriminative power in an earlier study.⁴² In the case of FCFP4 fingerprints, the distribution is centered around 0.4, which is also known as a good discriminative cutoff.⁴³ This maximum at a lower value is expected, as FCFP4 fingerprints are stricter in terms of Tanimoto score when dealing with bigger molecules. The fractions of dissimilar products obtained using two different fingerprints can be interpreted as hints that SCUBIDOO contains novel chemical entities in comparison with known drugs.

Synthetic Accessibility. To obtain a second computational assessment of the ease of synthesis of the products within SCUBIDOO, the SA score was computed for each product. The distribution of SA scores, plotted in Figure 8, is centered around of value of 3 with the vast majority (96%) lying below an SA score of 4, indicating easy-to-make products rather than overcomplex molecules.

Application: Retrospective Studies. In order to exemplify how SCUBIDOO might be useful in a ligand discovery context and to demonstrate several usage scenarios, retrospective similarity screening campaigns are presented here. The main goal is to show how SCUBIDOO can be used to retrieve known drugs or highly similar analogues. Table 1 lists all of the examples that were analyzed. The first example is meant to be comprehensive in order to demonstrate how to use SCUBIDOO step-by-step. Only examples 1, 2, and 3 are

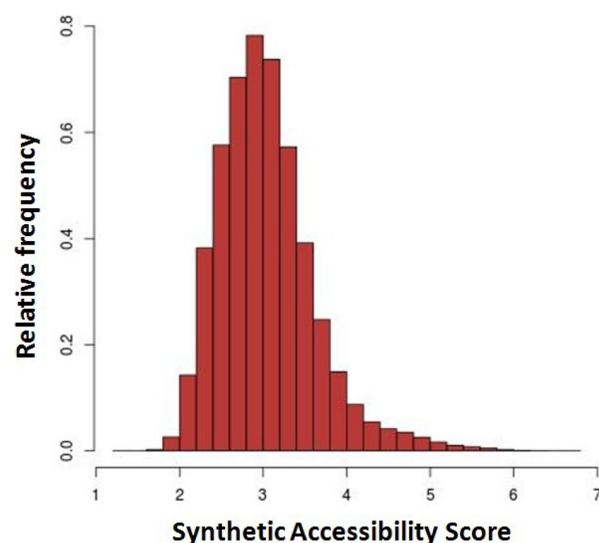


Figure 8. Distribution of SA scores for all of the products contained in SCUBIDOO.

described here. The remaining examples are provided in the Supporting Information.

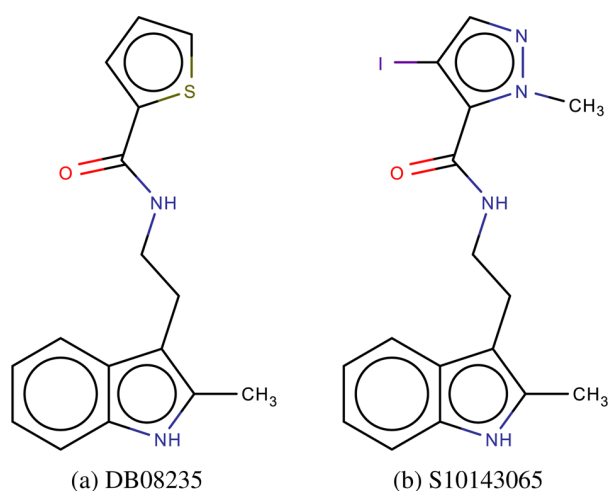
Example 1: DB08235. Ligand-Based Strategy. In the similarity screen between DrugBank and the S sample, a close match was identified between the experimental drug DB08235 and SCUBIDOO product S10143065 (Figure 9), with an FCFP4 Tanimoto score of 0.69. Both molecules have an amide group attached to an indole moiety. DB08235 is an experimental drug that was identified as an inhibitor of the Arp2/3 complex⁴⁴ and may be utilized as potential anticancer agent. A one-click search in SCUBIDOO's Web interface retrieved information for S10143065. It is predicted to be synthesizable using a Schotten–Baumann amide reaction between the two building blocks 4029192 and 4089476 (Scheme 1). Building block 4089476 is particularly interesting here, as it contains the indole moiety. We then looked for every product in SCUBIDOO made from building block 4089476 using a Schotten–Baumann amide reaction, since this reaction introduces an amide bond, and 4460 derivative products were retrieved and compared to DB08235 using FCFP4 fingerprints. This search identified five products with Tanimoto scores above 0.84 (Figure 10). Among those, two very close analogues of DB08235 were present. Indeed, products S00003866 and S00021706 contain an isoxazole moiety and a thiazole moiety, respectively, which are very close to the thiophene moiety of DB08235. This application shows that after a molecule in the S sample that is similar to a given drug is identified, we can use the synthetic information on the product of interest to quickly screen the entire SCUBIDOO data set using a building block identifier and a reaction. A full search of the entire library would have taken several hours, as opposed to the few minutes for screening of the S sample and the 4460 derivatives of the original “hit”. Therefore, we were able to efficiently analyze the analogues based on building block 4089476 and retrieve five products that are more similar to the drug than the initial hit found in the S sample. This example also illustrates how SCUBIDOO can be used for structure–activity relationship (SAR) studies or to generate suggestions for fragment-growing strategies.

Structure-Based Assessment of DB08235 Analogues. The five analogues of DB08235 were then docked into the Arp2/3

Table 1. Summary of the Hits Found within SCUBIDOO after Similarity Screening against DrugBank and the PDB Based on the FCFP4 Fingerprints

ex ^a	ref ID ^b	hit ID ^c	sim (FCFP4) ^d	reaction ^e	ref set ^f	sample ^g
1	DB08235	S00003866	0.96	amide	DrugBank	S
2	DB01097	S00131967	0.68	amide	DrugBank	M
3	H50	S02142952	0.93	Suzuki	PDB	M
3	H50	S02148982	0.94	Suzuki	PDB	M
4	4K6	S07366028	1	amide	PDB	S
5	F8E	S13393814	1	Buchwald–Hartwig	PDB	L
6	RM8	S01918821	1	amination	PDB	L
7	1DZ	S16929461	1	N-arylation	PDB	L

^aExample. ^bReference ID (drug or PDB ligand). ^cSCUBIDOO ID. ^dTanimoto similarity score using the FCFP4 fingerprints. ^ePredicted reaction. ^fReference data set. ^gSCUBIDOO sample.

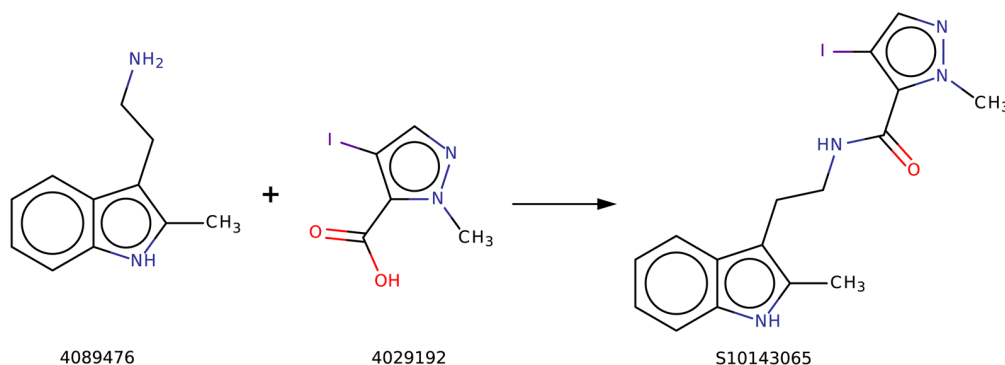
**Figure 9.** Molecules mentioned in example 1: (a) DrugBank compound DB08235 and (b) its closest similar product in the S sample, S10143065.

complex using FRED.⁴⁵ The crystal structure is available in the PDB (ID 3DXK). The predicted binding mode for product S00021706, illustrated in Figure 11, overlaps perfectly with the cocrystallized ligand DB08235. The same binding mode was also retrieved for the remaining four analogue products. This can be taken as a hint that the two close analogues might exhibit similar biological activities, as they scored favorably in this orthogonal screening method. Another way to look at this result is that if S00021706 had been suggested as a potential ligand in an unbiased docking screen, one would have been able

to readily retrieve the other analogues quickly, leading to potentially biologically active compounds.

Example 2: DB01097. After screening of the DrugBank against the M sample, a close match was identified between leflunomide (DB01097) and product S00134656 (Figure 12) with an FCFP4 Tanimoto score of 0.67. They both contain building block 3001678. Focusing on this building block, similar to the procedure of example 1, led to the identification of the closest product to leflunomide in the entire SCUBIDOO database, S00131967, with an FCFP4 Tanimoto score of 0.68. The only difference between S00131967 and leflunomide is that the benzene ring is replaced by a pyridine ring. However, the isoxazole ring, which is the active part and is opened upon administration,⁴⁶ is identical. S00131967 is predicted to be synthesizable using a Schotten–Baumann amide reaction. This reaction was also applied to synthesize leflunomide.⁴⁷

Example 3: H50. In the comparison of the ligands from the PDB against the M sample, a similarity appeared between the fibrillogenesis inhibitor H50⁴⁸ (Figure 13a) and product S03544112 with an FCFP4 Tanimoto score of 0.84 (Figure 13b). Product S03544112 is predicted to be synthesizable using a Suzuki coupling between the building blocks 4003301 and 6644827. Suzuki coupling was also applied to synthesize H50.⁴⁹ In this case, the two molecules do not share a common building block. Exploring the analogues of building block 4003301 led to the boronic-acid-containing building block 3200974, which is more similar to the initial H50, as only a chlorine is replaced by a fluorine. The derivatives of building block 3200974 obtainable by Suzuki coupling were then compared to H50, and product S02142952 (Figure 13c) was identified as a closer analogue, with an FCFP4 Tanimoto similarity score of 0.93. Furthermore,

Scheme 1. Route for Obtaining S10143065 (right): The Schotten–Baumann Amide Reaction between the Two Building Blocks 4089476 (left) and 4029192 (middle)

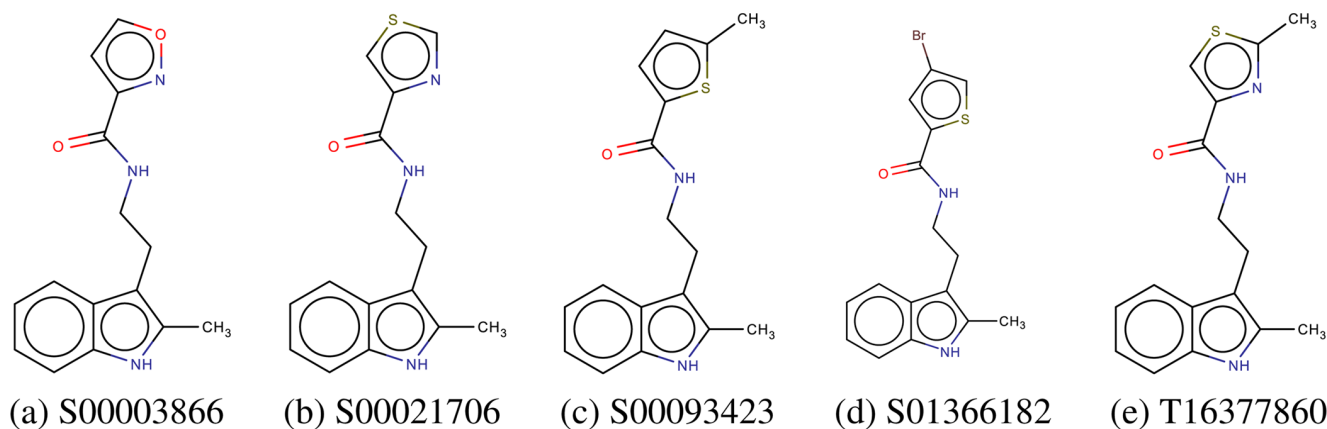


Figure 10. (a–e) Molecular analogues of DrugBank compound DB08235 found in SCUBIDOO.

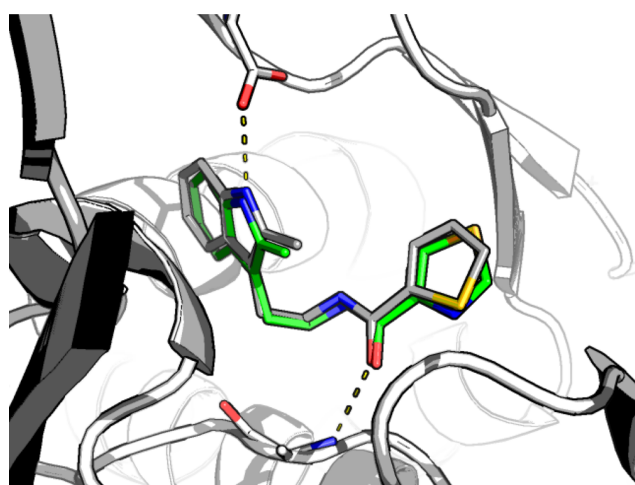


Figure 11. Docking: predicted binding mode of product S00021706 (green carbons) compared to the crystallized ligand DB08235 (gray carbons). The protein is shown in white cartoon representation, and H-bonds are indicated by yellow dashed lines.

the derivatives of building block 4003301 were compared to H50, leading to yet another close analogue, product S02148982 (Figure 13d), with a similarity score of 0.94.

DISCUSSION AND CONCLUSIONS

We have presented a freely accessible database concept currently holding 21 million screenable chemical products, each coming with synthesis information allowing an estimation of how readily it might be obtained. All of the reactants used for the creation of this database are publicly available, and the

reactions employed are among the most popular ones in the pharmaceutical field. The products are accessible using an intuitive Web interface, complete with synthesis instructions. SCUBIDOO is unique because it not only provides reaction suggestions but also clearly specifies potential side or alternative reactions and even reactive groups that could interfere during synthesis. Such warnings can help chemists as decision tools during synthesis planning and protecting group design. Moreover, since SCUBIDOO is a Web-based application, gathering of feedback from the scientific community is possible and will be applied to refine the products and reactions in future versions. This will benefit the community by making the set of reactions and the alerts in SCUBIDOO even more robust.¹⁷

It is clear that the diversity of the initial building block library has a large impact on the generated products. Therefore, future versions or derivative libraries will originate from new building blocks and increase the diversity. Conversely, in order to broaden the library more, it could be interesting to ensure that the building block library provides some heterogeneity within the reagent classes. This will enhance the diversity of the generated products and allow the use of the set of 58 reactions at its full potential. As a possible future development of their work,²³ Goldberg et al. suggested an open-innovation approach where ideas for novel structures to synthesize could be accessed from external sources. SCUBIDOO fits right into this context.

SCUBIDOO is not so much a database as it is a concept of how to enter chemical space. Along those lines, the PCA analysis suggested a share of novelties within SCUBIDOO that can be assumed to be “low-hanging fruits” as described by Hartenfeller et al.¹² More precisely, such chemical scaffolds undescribed for a particular target can be used as starting points

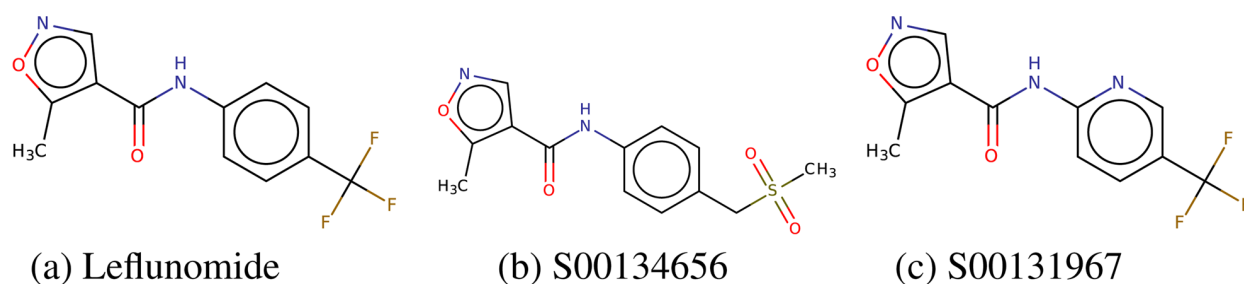


Figure 12. (a) DrugBank compound DB01097 (leflunomide). (b) Its closest similar product in the S sample, S00134656 (Tanimoto score = 0.67). (c) Its closest similar product in all of SCUBIDOO, S00131967 (Tanimoto score = 0.68).

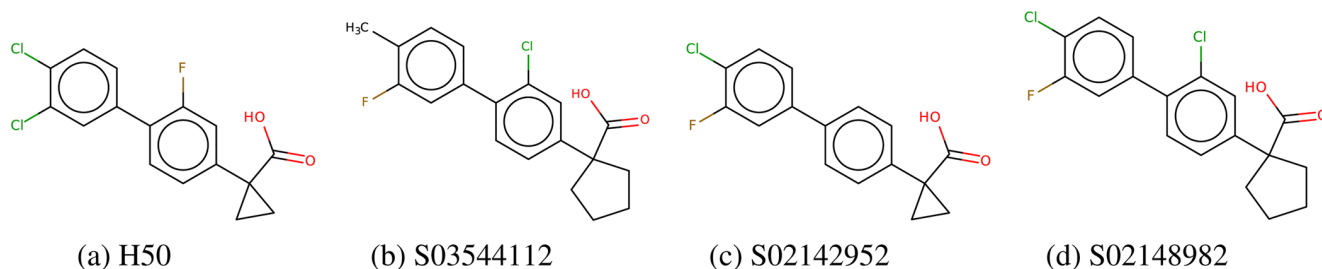


Figure 13. (a) Ligand H50. (b) Its closest similar product in the M sample, S03544112 (Tanimoto score = 0.84). (c) A similar product in SCUBIDOO, S02142952 (Tanimoto score = 0.93). (d) Its closest similar product in SCUBIDOO, S02148982 (Tanimoto score = 0.94).

for ligand discovery projects. Once such a “fruit” is identified, SCUBIDOO allows users to rapidly explore the tree around this point in order to harvest close analogues. This analogy is illustrated in the retrospective studies, where we identified existing active molecules within SCUBIDOO starting from the samples. Whenever possible, we retrieved information about the original synthesis, which matches the category of SCUBIDOO well. In a prospective screening setting, where such suggestions might come from docking, SCUBIDOO can be used in a straightforward manner to assemble a tailored library. Since this method also proceeds via a close look at the fragment composition, it fits right in with the ALTA approach, which we have described earlier.⁵⁰ Additionally, in cases where only analogues of active molecules were retrieved within the samples, a second, refined, search at the building block level allowed us to retrieve the initial active molecule or a very close analogue in all cases discussed here. Retrospective binding mode analysis showed that the small chemical differences should not affect the ligand–protein interactions.

While the size of this database is relatively small compared with those of other virtual chemical space libraries,¹⁵ navigating through 21 million entities already presents a challenge. In order to provide an entry point, SCUBIDOO was reduced to three different representative samples denoted as S, M, and L, containing 9994, 99 977, and 999 794 compounds, respectively. The representative samples were extracted using a stratified balanced sampling algorithm, which is a well-known algorithm for population surveys. Its application to a large molecule set allowed us to obtain samples respecting the heterogeneity of the initial set. This is essential in order to use the database in an efficient fashion in future screens. We are aware that 21 million products is far from covering chemical space. Nevertheless, we think that concomitant with future expansion of our database, such reduction steps need to be applied in order to keep the data computationally tractable. To the best of our knowledge, this is the first application of balanced sampling for this purpose.

A typical application protocol might start with screening of the SCUBIDOO sample of a user’s choice using a ligand-based strategy, a structure-based strategy, or both. The second step would then involve a focused search around the candidate hits found during the first screening, this time extracting molecules from the entire database.

Furthermore, SCUBIDOO can also be employed as a growing strategy within a fragment-based project or as an SAR tool. Indeed, any SCUBIDOO product is the assembly of two building blocks (or fragments). If one of those building blocks shows promising interactions with a given target, SCUBIDOO lets users quickly retrieve all of the derivatives of this building

block. All of these products can then be downloaded in SMILES format for further investigation.

While the retrospective assessment shows the existence of known active molecules within SCUBIDOO and the SA scores suggest that most of the products fall on the side of relatively facile synthetic realization, the next step will be to validate products experimentally. We think that this database has the potential to go in the direction of one of the expected breakthroughs in future de novo design, as stated by Schneider:⁵¹ “reliable prediction of the synthesizability of new chemical entities and suggestion of short synthesis routes and reactions, directly coupled to integrated synthesis-and-test platforms”. After the recent publication of the “synthesis machine”,⁵ it is not ludicrous to think that the first brick of such a workflow might be a database based on the SCUBIDOO concept.

■ ASSOCIATED CONTENT

● Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00203.

Spreadsheet containing the library of 58 reactions (encoded in SMARTS notation), the reagent classes, the nucleophile and electrophile groups (encoded as SMARTS), the 13 reactions that do not appear in SCUBIDOO, and PCA results on diversity and novelty (XLSX)

Discussion of retrospective study examples 4–7 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: peter.kolb@uni-marburg.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

P.K. thanks the German Research Foundation (DFG) for Emmy Noether Fellowship KO4095/1-1 and the LOEWE Program for participation in the research cluster SynChemBio. The authors acknowledge J. Cramer, Dr. J. Schiebel, Dr. A. Metz, and C. Taylor for their valuable input through discussions and reading of the manuscript. The authors also thank OpenEye Scientific Software for providing the software referenced in this study and permitting the generated products to be made publicly available.

■ REFERENCES

- (1) Lederberg, J. Topological mapping of organic molecules. *Proc. Natl. Acad. Sci. U. S. A.* **1965**, *53*, 134–139.

- (2) Fink, T.; Raymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discover. *J. Chem. Inf. Model.* **2007**, *47*, 342–353.
- (3) Reutlinger, M.; Rodrigues, T.; Schneider, P.; Schneider, G. Multi-objective molecular de novo design by adaptive fragment prioritization. *Angew. Chem., Int. Ed.* **2014**, *53*, 4244–4248.
- (4) Service, R. F. The Synthesis Machine. *Science* **2015**, *347*, 1190–1193.
- (5) Li, J.; Ballmer, S. G.; Gillis, E. P.; Fujii, S.; Schmidt, M. J.; Palazzolo, A. M. E.; Lehmann, J. W.; Morehouse, G. F.; Burke, M. D. Synthesis of many different types of organic small molecules using one automated process. *Science* **2015**, *347*, 1221–1226.
- (6) Alvim-Gaston, M.; Grese, T.; Mahoui, A.; Palkowitz, A. D.; Pineiro-Nunez, M.; Watson, I. Open Innovation Drug Discovery (OIDD): a potential path to novel therapeutic chemical space. *Curr. Top. Med. Chem.* **2014**, *14*, 294–303.
- (7) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Raymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.
- (8) Kolb, P.; Irwin, J. J. Docking screens: right for the right reasons? *Curr. Top. Med. Chem.* **2009**, *9*, 755–770.
- (9) Nikitin, S.; Zaitseva, N.; Demina, O.; Solovieva, V.; Mazin, E.; Mikhalev, S.; Smolov, M.; Rubinov, A.; Vlasov, P.; Lepikhin, D.; Khachko, D.; Fokin, V.; Queen, C.; Zosimov, V. A very large diversity space of synthetically accessible compounds for use with drug design programs. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 47–63.
- (10) Cramer, R. D.; Soltanshahi, F.; Jilek, R.; Campbell, B. AllChem: Generating and searching 1020 synthetically accessible structures. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 341–350.
- (11) Rarey, M.; Stahl, M. Similarity searching in large combinatorial chemistry spaces. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 497–520.
- (12) Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K. H.; Schneider, G.; Jacoby, E.; Renner, S. Probing the bioactivity-relevant chemical space of robust reactions and common molecular building blocks. *J. Chem. Inf. Model.* **2012**, *52*, 1167–1178.
- (13) Lessel, U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H. Searching fragment spaces with feature trees. *J. Chem. Inf. Model.* **2009**, *49*, 270–279.
- (14) Hu, Q.; Peng, Z.; Kostrowicki, J.; Kuki, A. Methods in Molecular Biology. *Chemical Library Design* **2011**, *685*, 253–276.
- (15) Peng, Z. Very large virtual compound spaces: Construction, storage and utility in drug discovery. *Drug Discovery Today: Technol.* **2013**, *10*, e387–e394.
- (16) Gasteiger, J. Cheminformatics: Computing Target Complexity. *Nat. Chem.* **2015**, *7*, 619–620.
- (17) Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K.-H.; Schneider, G.; Jacoby, E.; Renner, S. A Collection of Robust Organic Synthesis Reactions for In Silico Molecule Design. *J. Chem. Inf. Model.* **2011**, *51*, 3093–3098.
- (18) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, *54*, 3451–3479.
- (19) Hann, M. M. Molecular obesity, potency and other additions in drug discovery. *MedChemComm* **2011**, *2*, 349–355.
- (20) ChemBridge. Building Blocks. http://www.chembridge.com/building_blocks/.
- (21) Landrum, G. RDKit: Open-source chemoinformatics. <http://www.rdkit.org>.
- (22) Cecchini, M.; Kolb, P.; Majeux, N.; Caffisch, A. Automated docking of highly flexible ligands by genetic algorithms: a critical assessment. *J. Comput. Chem.* **2004**, *25*, 412–422.
- (23) Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W. D.; Tomkinson, N. P. Designing novel building blocks is an overlooked strategy to improve compound quality. *Drug Discovery Today* **2015**, *20*, 11–17.
- (24) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740.
- (25) OMEGA, version 2.5.1.4; OpenEye Scientific Software: Santa Fe, NM; <http://www.eyesopen.com>.
- (26) Chauvet, G. Stratified balanced sampling. *Surv. Methodol.* **2009**, *35*, 115–119.
- (27) Grafstrom, A. BalancedSampling: Balanced and spatially balanced sampling. In R package version 1.4, 2014.
- (28) R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2008.
- (29) Deville, J.-c.; Tillé, Y. Efficient balanced sampling: The cube method. *Biometrika* **2004**, *91*, 893–912.
- (30) Hasler, C.; Tillé, Y. Fast balanced sampling for highly stratified population. *Comput. Stat. Data Anal.* **2014**, *74*, 81–94.
- (31) DrugBank: Open Data Drug & Drug Target Database, version 4.1; <http://www.drugbank.ca/>.
- (32) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
- (33) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (34) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and cambridge structural database. *J. Chem. Inf. Model.* **2010**, *50*, 572–584.
- (35) QUACPAC, version 1.6.3.1; OpenEye Scientific Software: Santa Fe, NM; <http://www.eyesopen.com>.
- (36) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (37) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8.
- (38) Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599–3601.
- (39) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Model.* **1999**, *39*, 868–873.
- (40) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714–3717.
- (41) McGregor, M. J.; Pallai, P. V. Clustering of Large Database of Compounds: Using MDL keys As Structural Descriptors. *J. Chem. Inf. Model.* **1997**, *37*, 443–448.
- (42) Chevillard, F.; Lagorce, D.; Reynès, C.; Villoutreix, B. O.; Vayer, P.; Miteva, M. Multimodel Protocol Based on Chemical Similarity In silico Prediction of Aqueous Solubility: A Multimodel Protocol Based on Chemical Similarity. *Mol. Pharmaceutics* **2012**, *9*, 3127–3135.
- (43) Wawer, M.; Bajorath, J. Similarity-potency trees: A method to search for SAR information in compound data sets and derive SAR rules. *J. Chem. Inf. Model.* **2010**, *50*, 1395–1409.
- (44) Nolen, B. J.; Tomasevic, N.; Russell, A.; Pierce, D. W.; Jia, Z.; McCormick, C. D.; Hartman, J.; Sakowicz, R.; Pollard, T. D. Characterization of two classes of small molecule inhibitors of Arp2/3 complex. *Nature* **2009**, *460*, 1031–1034.
- (45) McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2011**, *51*, 578–596.
- (46) Liu, S.; Neidhardt, E.; Grossman, T.; Ocain, T.; Clardy, J. Structures of human dihydroorotate dehydrogenase in complex with antiproliferative agents. *Structure* **2000**, *8*, 25–33.
- (47) Ivashkin, P.; Lemonnier, G.; Cousin, J.; Grégoire, V.; Labar, D.; Jubault, P.; Pannecoucke, X. [¹⁸F]CuCF₃: A [¹⁸F]Trifluoromethylating Agent for Arylboronic Acids and Aryl Iodides. *Chem. - Eur. J.* **2014**, *20*, 9514–9518.
- (48) Zanotti, G.; Cendron, L.; Folli, C.; Florio, P.; Imbimbo, B. P.; Berni, R. Structural evidence for native state stabilization of a conformationally labile amyloidogenic transthyretin variant by fibrillogenesis inhibitors. *FEBS Lett.* **2013**, *587*, 2325–2331.

(49) Peretto, I.; et al. Synthesis and biological activity of flurbiprofen analogues as selective inhibitors of beta-amyloid42 secretion. *J. Med. Chem.* **2005**, *48*, 5705–20.

(50) Kolb, P.; Berset Kipouros, C.; Huang, D.; Caflisch, A. Structure-based tailoring of compound libraries for high-throughput screening: Discovery of novel EphB4 kinase inhibitors. *Proteins: Struct., Funct., Genet.* **2008**, *73*, 11–18.

(51) Schneider, G. Future de novo drug design. *Mol. Inf.* **2014**, *33*, 397–402.

5.2 Supporting Information

SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward high Likelihood of Synthetic Tractability

Supporting Information

Chevillard F. and Kolb P.*

Department of Pharmaceutical Chemistry, Philipps-University Marburg, Germany

E-mail: peter.kolb@uni-marburg.de

Results

Application: retrospective studies

In this section the examples which are not fully described in the paper are presented (cf Table 1).

Example 4: 4K6

Screening the ligands from the PDB against the S sample allowed to show that 4K6 is present in this sample under the ID S7366028 (Figure S1). 4K6 is an inhibitor of factor IXa which was recently crystallized. Product S7366028 is predicted to be synthesizable through an amide formation, which is also the case for 4K6.¹

*To whom correspondence should be addressed

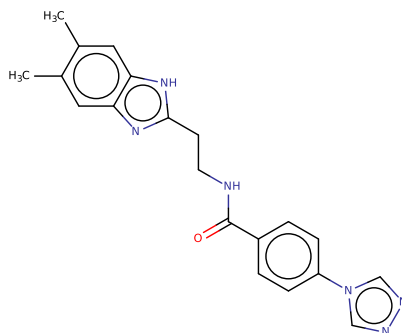


Figure S1: 4K6

Example 5: F8E

Matching the ligands from the PDB to their close counterparts in the L sample yielded a close match between F8E,² which is a Tau-Tubulin Kinase 1 inhibitor, and product S13392303 (FCFP4 Tanimoto = 0.91). These molecules share the same building block 4036676. After comparing the derivatives of 4036676 reactive in a Buchwald-Hartwig reaction, we were able to retrieve product S13393814 (FCFP4 Tanimoto = 1), which only differs from F8E by a substitution between a chlorine and a bromine (Figure S2). Product S13393814 was then docked within Tau-tubulin kinase 1 receptor using GOLD.³ Two water molecules occupy the binding site and were kept for the docking process. The binding mode found for product S13393814 overlaps with the crystallized F8E and all the polar interactions were recapitulated (Figure S3).

Example 6: RM8

Screening the ligands from the PDB against the L sample yielded another close match between RM8,⁴ which is an acidic mammalian chitinase inhibitor, and product S03114005 (FCFP4 Tanimoto = 0.84). These molecules share the same building block 3020388. Derivatives of building block 3020388 pointed us towards RM8 itself, present under the ID S01918821 (FCFP4 Tanimoto = 1).

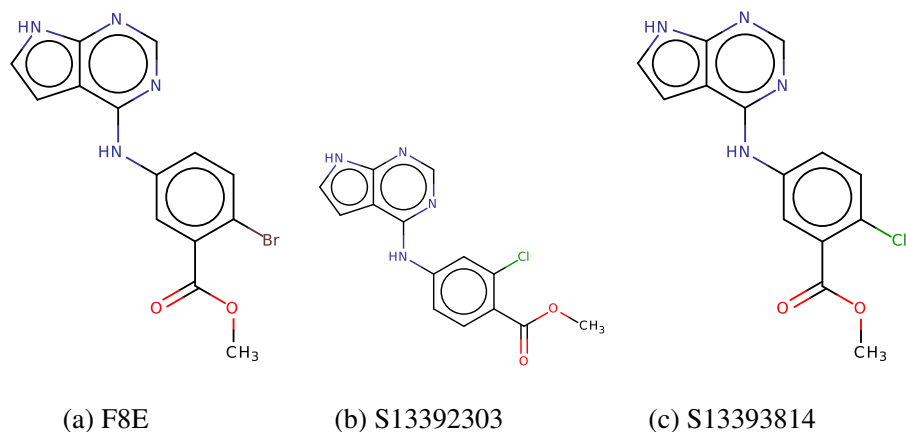


Figure S2: (a) F8E. (b) Its closest similar product S13392303 in the L sample (FCFP4 Tanimoto = 0.91). (c) Its closest similar product S13393814 in SCUBIDOO (FCFP4 Tanimoto = 1).

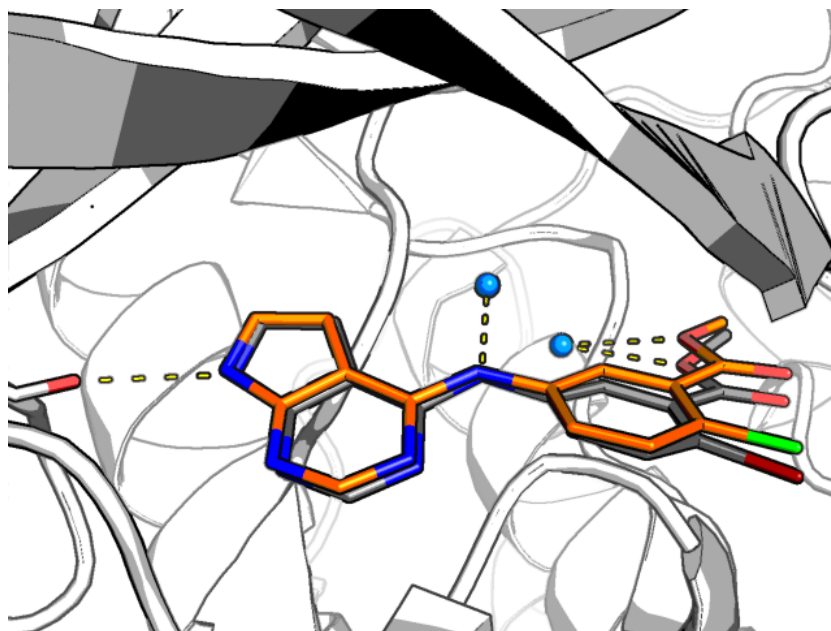


Figure S3: Docking: predicted binding mode of product S13393814 (orange carbons) compared to the crystallized ligand F8E (grey carbons). Protein in white cartoon representation. H-bonds are indicated by yellow dashed lines.

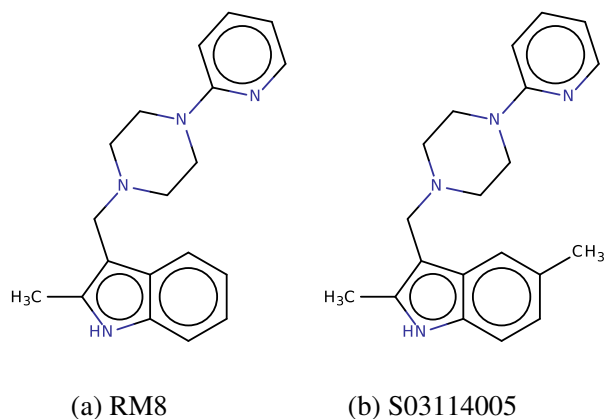


Figure S4: (a) RM8. (b) Its closest similar product S03114005 in the L sample (FCFP4 Tanimoto = 0.90).

Example 7: 1DZ

Finally, the comparison of the PDB against the L sample identified 1DZ, which is a PPI inhibitor,⁵ and product S16936876 as a close match (FCFP4 Tanimoto = 0.83). These two molecules do not share a common building block. However, exploring the analogs of building block 7278181 led to building block 7253309, contained in the initial 1DZ. The derivatives of building block 7253309 amenable to N-arylation on a heterocycle were then screened against 1DZ and 1DZ was retrieved under the ID S16929461.

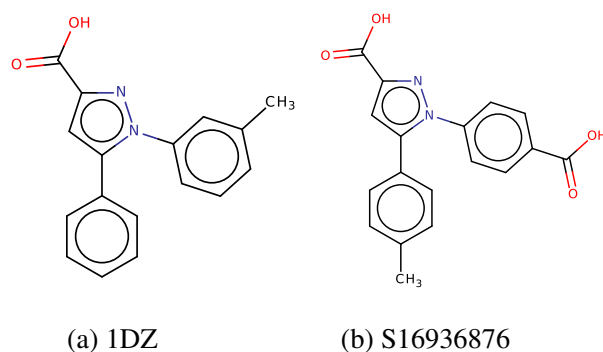


Figure S5: (a) 1DZ. (b) Its closest similar product S16936876 in the L sample (Tanimoto = 0.83).

References

- (1) Parker, D. L. et al. *Bioorganic & Medicinal Chemistry Letters* **2015**, 25, 2321–2325.
- (2) Xue, Y.; Wan, P. T.; Hillertz, P.; Schweikart, F.; Zhao, Y.; Wissler, L.; Dekker, N. *ChemMedChem* **2013**, 8, 1846–1854.
- (3) Jones, G.; Willett, P.; Glen, R. C. *J. Mol. Biol.* **1995**, 245, 43–53.
- (4) Cole, D. C. et al. *J. Med. Chem.* **2010**, 53, 6122–6128.
- (5) Frank, A. O.; Feldkamp, M. D.; Kennedy, J. P.; Waterson, A. G.; Pelz, N. F.; Patrone, J. D.; Vangamudi, B.; Camper, D. V.; Rossanese, O. W.; Chazin, W. J.; Fesik, S. W. *J. Med. Chem.* **2013**, 56, 9242–9250.

"Opportunities multiply as they are seized"

Sun Tzu

The chapter 6 is a scaffold of an article. I was responsible of the overall strategy, ligand based approach, docking screening with OpenEye tools, visual inspection and selection of the fragments to optimize. Christof Siefker handled the TSA and the crystallization of fragment 4012413. Helena Rimmer synthesized the fragment 4012413 and is currently working on derived products. Corey Taylor helped with the docking screening with DOCK (results not shown here) and evaluation of the poses from FRED screening.

Chapter 6

Design and identification of novel ligands for the PIM1 kinase

6.1 Introduction

Proviral integration Maloney (PIM) kinases belong to the family of protein serine / threonine kinases and consists of three members (PIM1, PIM2 and PIM3). PIM kinases are involved in cell proliferation and survival, and has been shown to be overexpressed in a variety of tumors, namely pancreatic, prostate and colon [262–269]. PIM1 is mainly expressed in hematopoietic cells [270] and has been shown to be overexpressed in a wide range of human leukemias. Thus PIM kinases are particularly important therapeutic targets in oncology [271].

Given the relevance of these targets, we began a fragment-based ligand discovery campaign focusing on PIM1 using computational approaches as first screening technique. A recent study has shown that *de novo* fragment growing strategies were successfully applied in the identification of low nanomolar PIM kinase inhibitors [271]. The authors designed their two initial compounds with previously published structures as starting points [272]. They assumed that the compounds were able to interact with the catalytic Lys67 and also hypothesized that they might be involved in a salt bridge with Asp128 and/or Glu171. Docking predictions suggested that the *S*-isomers made direct interactions to Asp128 and Glu171, while the *R*-isomers did not. Assays revealed that the *S*-isomers were 5-20-fold more active. Thus the authors suggested that a direct binding to Asp128 and/or Glu171 could be necessary to achieve a higher affinity. Later on, crystal structures revealed that interactions with Lys67 were correctly predicted but one of the compounds did not make direct interactions to Asp128 and Glu171. Instead the

compound made mediated interactions to those residues via a water molecule, highlighting that a direct interaction to Asp128 and Glu171 was not strictly necessary to achieve a leap in affinity.

In our study, we applied a two-step computational strategy in order to design novel ligands for the PIM1 kinase. The first step consists of identifying fragments which make compelling interactions with Glu121 and/or Lys67. The second step consist of growing the fragment into a larger product that could interact with Asp128 and/or Glu171. To do so we made use of structure-based and ligand-based approaches in synergy with the SCUBIDOO database [261]. This database currently holds 21 M virtual products that are the assembly of two available building blocks. Products were created using a collection of robust organic reactions [1, 2]. This *bipartite product* philosophy fits nicely with our strategy, where a first building block needs to entertain favorable interactions with Glu121 and/or Lys67, while the second one should interact with Asp128 and/or Glu171.

This study will first presents how to use SCUBIDOO and provide to the scientific community a detailed guide of different scenarios. Three fragments were identified for further optimization, each using a different strategy. One was found using ligand-based approaches (similarity search), one was identified using structure-based techniques (docking) and the last one was discovered when trying to improve the hit from docking (analog search). Each of the three fragments was then grown using a particular chemical reaction. Doing so, allows one to help validating the synthetic feasibility of SCUBIDOO products by using different reactions. Ensuring chemical reaction diversity allows us to synthesize diverse compounds which should increase our chance to identify PIM1 inhibitors. Finally, we will highlight how one can make use of *in silico* screening in order to identify fragments with favorable ligand efficiency (LE). This will be illustrated with our ongoing results where one of the three initial fragment, identified through docking, was tested in a thermal shift assay (TSA) and showed a positive shift. This first experimental clue, was quickly followed by a second clue, a successful attempt to crystallize the fragment with PIM1. The structure of the complex should be solved in the near future (August 2016). If the predicted binding mode aligns with the crystal structure, this will confirm that the fragment is a hit and that we could optimized into a better ligand for the PIM1.

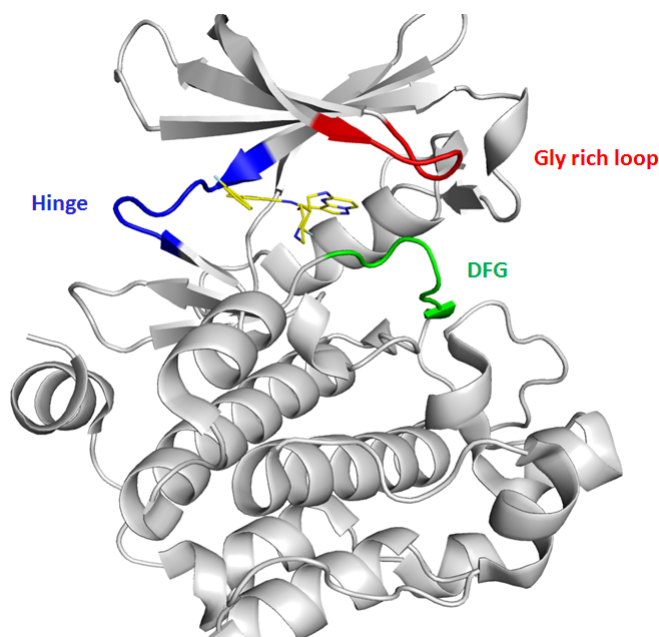


FIGURE 6.1: PIM1 crystal structure. The hinge region is colored in blue, the DFG motif in green and the glycine rich loop in red.

6.2 Methods

6.2.1 The PIM1 binding site

The PIM1 binding site consists of three main regions highly conserved among the kinase family: the hinge region, a glycine-rich loop and the DFG motif (Figure 6.1). The binding site can be regarded as a cave, with the “ceiling” (glycine rich loop) and the “floor” being mostly hydrophobic. In between stands the hinge region containing the conserved Glu121 which is known to interact with ATP via an H-bond between the backbone carbonyl and the adenine moiety. The catalytic Lys67 is an important residue which has been shown to be involved in the inhibition of PIM1 [273–276]. Some inhibitors are also known to form a salt bridge with Asp128 or Glu171 [275–278] which is located below the hinge, in a more solvent-exposed region (Figure 6.2).

6.2.2 Bi-partite product philosophy

Any product in SCUBIDOO is the assembly of two building blocks (chapter 5). Thus any product can be divided in two fragments A and B . For the remaining applications of SCUBIDOO, A will be defined as the key fragment, the one that makes the most compelling interactions and thus the fragment we want to optimize (growing). B will be the fragment attached (extension) to A that can be assimilated as additional steric or polar constraints (the less important part of the product).

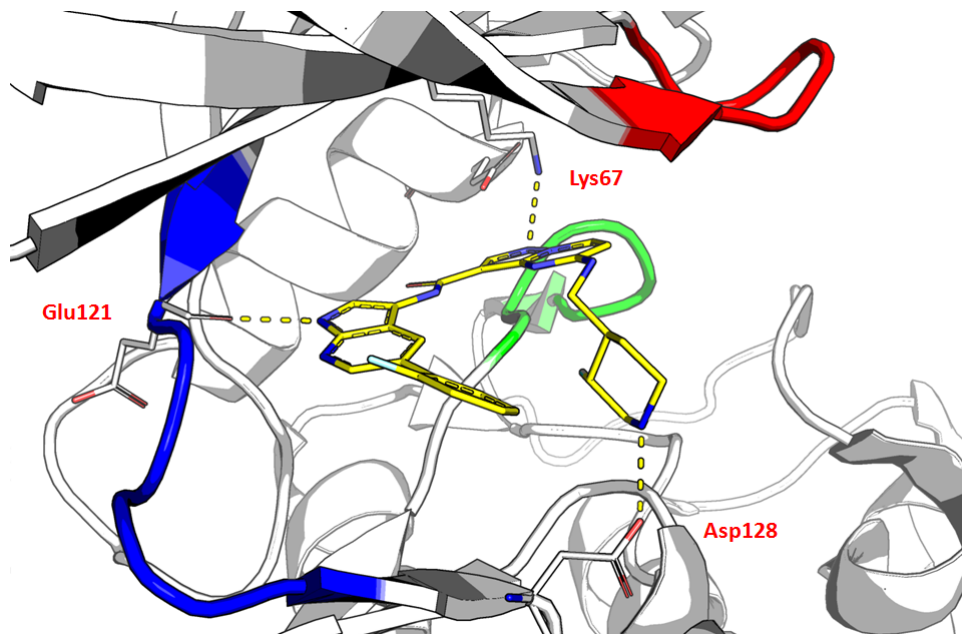


FIGURE 6.2: PIM1 key residues targeted by our growing strategy.

6.2.3 Growing strategy

Structure-based and ligand-based strategies were applied in order to identify *A* fragments that could interact with at least the conserved carbonyl of Glu121 or the catalytic Lys67. A perfect scenario would be interactions with both residues. For further consideration, *A* fragments should also contains reactive features (i.e. chemical features compatible with the organic reactions) pointing towards the solvent so one can attach *B* extensions that are likely to interact with Asp128 or Glu171.

6.2.4 Receptor preparation

Docking calculations were performed with the “DFG in” conformation (active) of PIM1 in complex with two high affinity inhibitors (PDB: 3BGP and 3VBV) [270, 279], the latter inhibitor being used in our fragment growing strategy. All ligands and solvent molecules were removed. The hydrogens were placed and minimized using the HBUILD module in CHARMM [249].

6.2.5 Active and decoy sets

All the molecules with a reported activity against PIM1 were downloaded from ChEMBL [280] and any molecule with a measured K_i lower than 10 μM was selected as active. This procedure yielded 730 active compounds. The decoys were generated using the DUD-E [281] web service, yielding 47’092 decoys.

6.2.6 Libraries preparation

The active and decoy sets, the S sample and all the derivative libraries (i.e. the derived products from a same building block) mentioned in this study were prepared for docking according to the same protocol:

- Compounds were protonated at physiological pH ($\text{pH} = 7.4$) and the most likely tautomer was assigned using QUACPAC [246].
- In case any protonated tertiary amines were present, both enantiomers were generated using *flipper* [254].
- Up to 500 conformers were generated for each molecule by means of OMEGA [57].

6.2.7 Docking

All products from the SCUBIDOO samples and the derivative libraries were docked with FRED [58, 80–82] which is described in chapter 2. The exceptions are the derived products of building block 5175110, which were docked using HYBRID. HYBRID shares the same algorithm with FRED with the exception that a component in the scoring function takes into account the 3D overlap with a reference crystal ligand. This feature is extremely useful in a fragment growing context, since products overlapping with the core fragment (i.e. the *A* fragment) will be scored higher than non overlapping one. SZYBKI [87] was used as a refinement approach in order to minimize the poses found and remove clashes (internal or external).

6.2.8 Chemical descriptors

Chemical similarity was based on the Tanimoto score of the FCFP4 fingerprints [30], which were computed using the RDKit library [257].

6.3 Results

6.3.1 Evaluation of PIM1 structures by means of enrichment calculations

The 102 crystal structures of PIM1 available in the protein data bank (PDB accessed in April 2015) were visually inspected in order to select a pool of diverse structures. This

was done by choosing crystal ligands that offer chemical diversity, thus different binding modes and therefore interact with different regions of the binding site. Ten structures were selected for further evaluation using enrichment calculations. This procedure aims at evaluating the discrimination power of a receptor by docking a set of known active ligands and a set of decoys (i.e. non-active). The more the receptor is able to retrieve active ligands in the top ranked molecules, the higher the enrichment, and thus the more reliable the receptor should be.

The enrichment results are summarized in table 6.1. Overall the results suggest a strong discriminative power ($AUC > 0.69$) for 9 receptors out of 10. Only 2XIY did perform relatively poorly with an AUC of 0.639. A comparison of 2XIY and 3BGP ($AUC = 0.814$) binding sites is illustrated in figure 6.3. Interestingly, in comparison to 3BGP, Asp186 in 2XIY is shifted towards Lys67 in order to accommodate an H-bond. This bond obstructs Lys67 in comparison to 3BGP, decreasing its polar surface area, thus increasing the polar constraints for any ligand to form an H-bond with Lys67. This could explain why the enrichment results are less robust with 2XIY, because it is harder for the ligands to interact with the catalytic Lys67, which is known to be crucial in the binding of many ligands [273–276].

Receptor (PDB)	AUC
3BGP	0.814
3F2A	0.794
4K18	0.792
4DTK	0.752
4K1B	0.752
3VBV	0.747
3C4E	0.742
2BIK	0.699
2J2I	0.690
2XIY	0.639

TABLE 6.1: Enrichment of a diverse set of PIM1 structures.

6.3.2 Ligand-based approach

The S sample and the Chembridge building blocks library (i.e. the library used to create SCUBIDOO) were screened against the active set using the FCFP4 fingerprints in order to identify products that were similar to known active ligands. This procedure yielded building block 5175110, which happens to be a crystallized ligand (PDB code = 0FK, 3VBV). A close inspection of 0FK’s binding mode (Figure 6.4) revealed a phenol moiety pointing towards the solvent and interacting with a water molecule. Derived products of

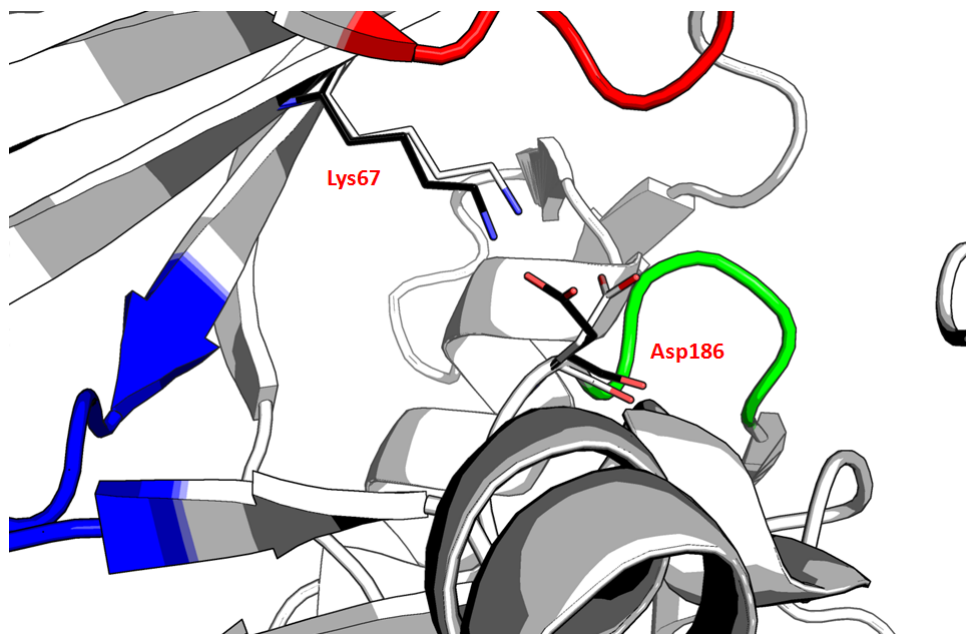


FIGURE 6.3: Comparison of PIM1 crystal structures 2XIY (black carbons) and 3BGP (white carbons).

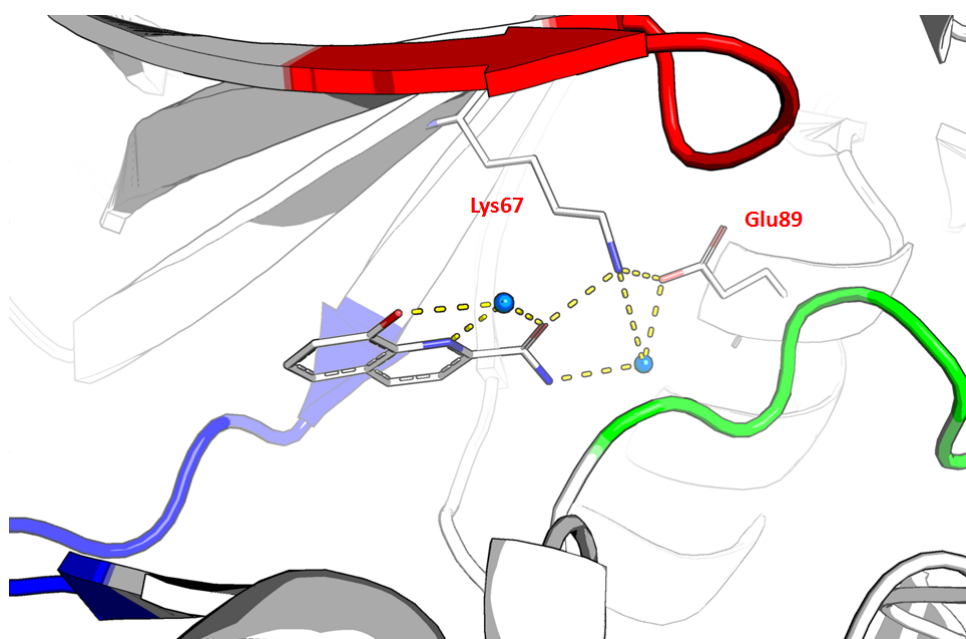


FIGURE 6.4: Fragment OFK (green carbons) crystallized in the PIM1 binding site (PDB 3VBV). Water molecules are represented in blue spheres and H-bond are represented as dashed yellow lines.

5175110 suggested that the phenol could be used for further growing with a *Mitsunobu phenol* reaction. Thus this building block was selected for further growing.

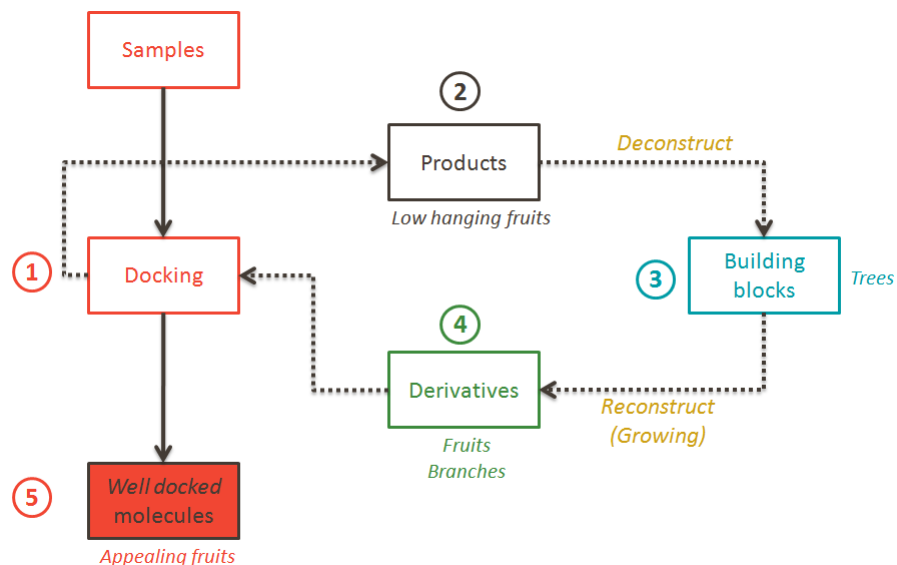


FIGURE 6.5: SCUBIDOO workflow for docking application.

6.3.3 Structure-based approach

A typical workflow for the use of SCUBIDOO in structure-based ligand discovery efforts is illustrated in figure 6.5. Each step will be detailed below.

6.3.3.1 Step 1: sample docking.

The S sample was docked in the binding site of PIM1 using FRED. The top 500 molecules based on the *score* were visually inspected, in order to identify *low-hanging fruits* (i.e. products containing building blocks that make compelling interactions with the protein).

6.3.3.2 Step 2: identification of low-hanging fruits.

Compound 3178025 is a typical *low-hanging fruit* (Figure 6.6) containing both a building block making compelling interactions and a building block which do not. The compelling building block (fragment *A*) contains a triazole moiety engaging the catalytic Lys67 via a H-bond and it also contains a phenylamine moiety which engages in favorable interactions with the carbonyl of Glu121. The second building block (fragment *B*) contains a protonated amine interacting too closely with Asp128 (distance N...O = 2 Å) thus clashing with the receptor. It also contains a free H-bond donor group (pyrrole) which is not involved in any interactions, thus penalizing the resulting compound even more.

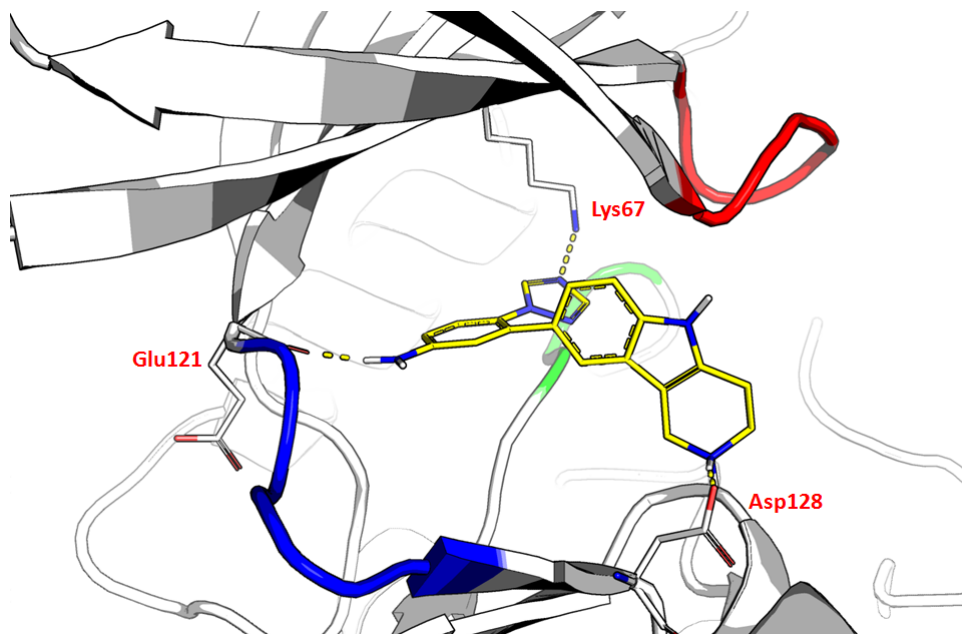


FIGURE 6.6: Low-hanging fruit (product 3178025) identified in the virtual screening of the PIM1 with the SCUBIDOO S sample.

6.3.3.3 Step 3: deconstruction of the low-hanging fruits.

Compound 3178025 was deconstructed in order to identify the reaction and building blocks necessary for its predicted synthesis. It was predicted to be synthesized using building blocks 4002721 and 4012414 via *Negishi coupling*. Building block 4012414 was selected for the next step, due to its nearly optimized interactions with the conserved residues of the binding site.

6.3.3.4 Step 4: construction of the building block derivatives.

The derived products of building block 4012414 based on all reactions compatible with the aryl halide were downloaded from SCUBIDOO. Five different reactions were thus investigated: *Grignard alcohol*, *Grignard carbonyl*, *Suzuki coupling*, *Negishi*, *Buchwald-Hartwig*.

6.3.3.5 Step 5a: docking derivatives and selection of the chemical reaction for growing.

The derived products for each of the five reaction (i.e. five series) were then docked in the binding site of PIM1. This procedure aimed at selecting which of the five reactions is compatible with our growing strategy. This selection was based on the evaluation of the binding mode of the derived products (i.e. can they interact with Asp128), which

was done visually by inspecting the top 500 molecules based on the *score*. Each reaction derivatives will be discussed briefly below.

Suzuki derivatives were too few (97) and were quickly discarded from this study. We took this decision because none of the generated products were contained a H-bond donor group which could interact with Asp128. Nonetheless, this allow us to underline that the library of boronic acids used to create SCUBIDOO is still to poor (only 97 building blocks) and stressed out the importance to repopulate this chemical family in future versions of SCUBIDOO.

Negishi derivatives were numerous (1514) and showed compelling products able to interact with Asp128. Thus this reaction was kept for our growing strategy.

Buchwald-Hartwig derivatives were plentiful (4662) due to the fact that building block 4012414 contains two reactive features compatible with this reaction (i.e. the aryl halide and the phenylamine). Only derivatives reacting with the aryl halide were kept, because they suggested a likely binding mode with Asp128. More interestingly, the substitution of the chlorine with an amine moiety introduces an H-bond donor group which seems able to form an internal H-bond with the triazole moiety, thus optimizing the polar surface area of 4012414. For all these reasons, this reaction was selected for our growing strategy.

Grignard alcohol derivatives contain a hydroxy group pointing towards an hydrophobic region of the receptor, thus introducing a free H-bond donor which will penalize the resulting product due to high desolvation cost that are not compensated upon binding. The same logic applies to *Grignard carbonyl* derivatives which increased the polar surface area with the introduction of a carbonyl moiety. For these reasons, those reactions were left out from our growing strategy.

6.3.3.6 Step 5b: optimization of the fragment to grow.

While the binding mode of 4012414 suggests that it will interact with both Lys67 and Glu121, the PSA of the fragment is not fully optimized. Indeed, only one hydrogen from the amine moiety is involved in a H-bond with the backbone carbonyl of Glu121, while the remaining hydrogen points towards a hydrophobic region of the receptor. This free hydrogen is likely to be desolvated without compensation and thus induces an enthalpy penalty for the overall product. Furthermore, for the *Negishi* derivatives, the triazole moiety contains a nitrogen which is not involved in any polar interactions. In order to remove this 'unnecessary' nitrogen we looked for analogs of 4012414 within the Chembridge library and identified 4012413, which answered our problem perfectly: the

triazole moiety was replaced by a diazole group. Thus, in order to start our growing strategy from pre-optimized fragments, we decided that the *Negishi* derivatives would be based on 4012413 and the *Buchwald-Hartwig* derivatives would be grown from 4012414.

6.3.3.7 Step 5c: selection of the products for synthesis

The derived products for the three *A* fragments 4012413, 4012414 and 5175110 (figure 6.9) were then docked in the binding site of the PIM1. The selection of the final products was based on the aptitude of satisfying a consistent binding mode for a given fragment *A*, as well as the faculty to interact with Asp128 or Glu171. This procedure was done visually by inspecting the top 500 molecules based on the *score* and repeated for each fragment *A*.

- 5175110: nine products were proposed for synthesis using *Mitsunobu phenol* reaction and all of them contained a protonated amine that is predicted to interact with Asp128. A suggestion for synthesis is illustrated in figure 6.7, where it can be seen that the pose of the suggested product overlaps nicely with the crystal ligand (i.e. OFK), thus preserving its initial binding mode.
- 4012414: 25 products were suggested for further synthesis using *Negishi* reaction, as they were deemed likely to interact with Asp128 or Glu171.
- 4012413: eight products were proposed for synthesis. This series contains only molecules predicted to form an internal H-bond between the triazole moiety and the amine group introduced by the *Buchwald-Hartwig* reaction.

6.4 Discussion

We have introduced the first application of SCUBIDOO in a ligand discovery effort focusing on PIM1. This work lays the foundations of how to use SCUBIDOO and is also an opportunity to learn about the database concept. Both ligand-based and structure-based approaches were used in concert in order to identify three fragments to grow and quickly design derived products for synthesis.

A ligand-based screening was applied in order to identify whether some known actives were similar to the building blocks used to create the database. The similarity search allowed us to identify a perfect match between building block 5175110 and a crystallized ligand (OFK). This building block was then computationally grown using the crystallized ligand as reference and by making use of the derived products present in SCUBIDOO (i.e.

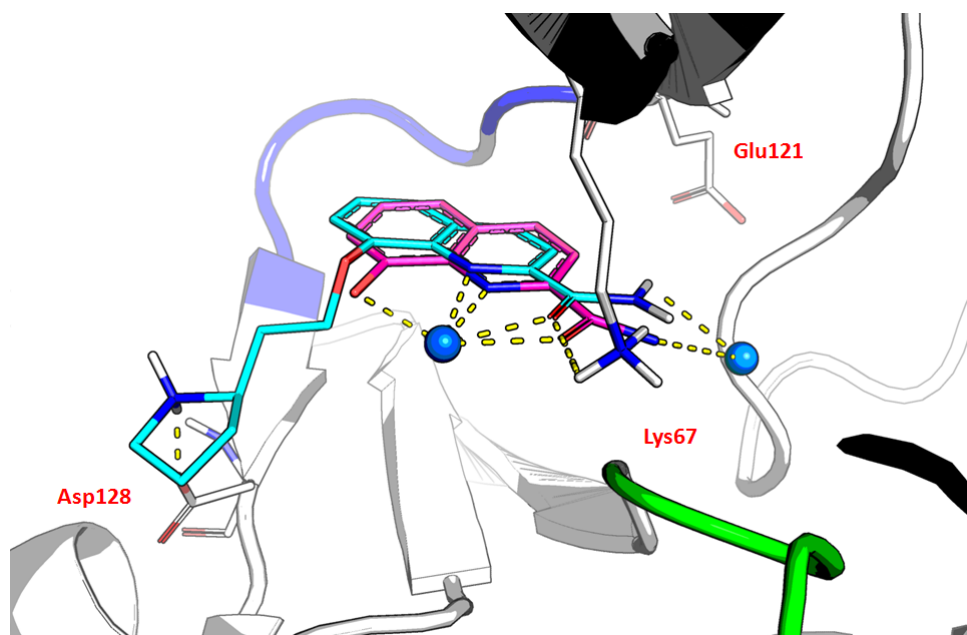


FIGURE 6.7: Illustration of a derived product of 5175110 (cyan carbons) using the *Mitsunobu phenol* reaction. The crystallized ligand (OFK) is colored with magenta carbons. Water molecules are represented as blue spheres and H-bond are represented with yellow dashed lines.

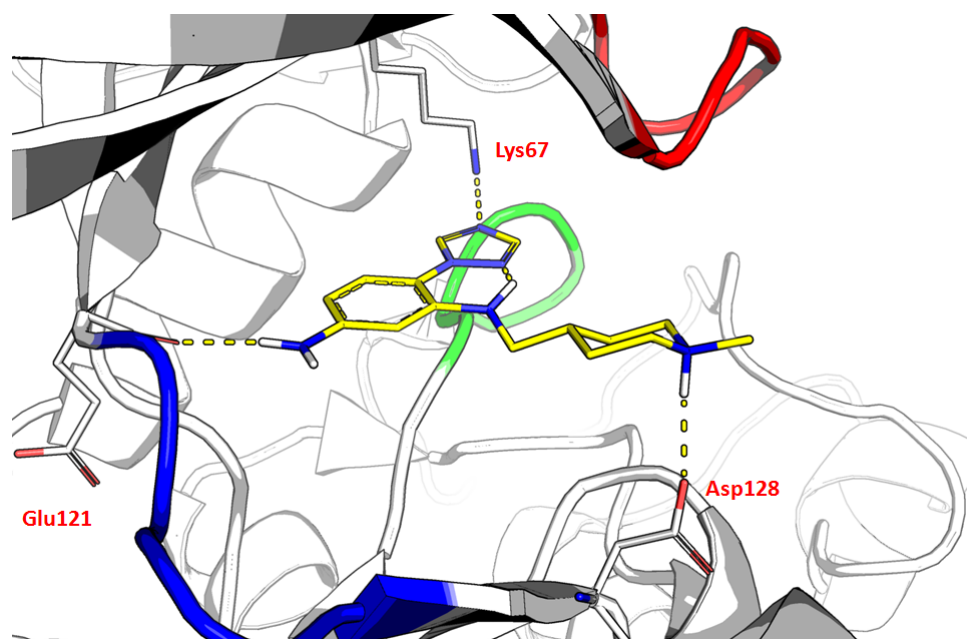


FIGURE 6.8: Illustration of a derived product from 4012414 (yellow carbons) using the *Buchwald-Hartwig* reaction. The generated product contains an amine moiety which might form an intramolecular H-bond with the triazole. H-bond are represented with yellow dashed lines.

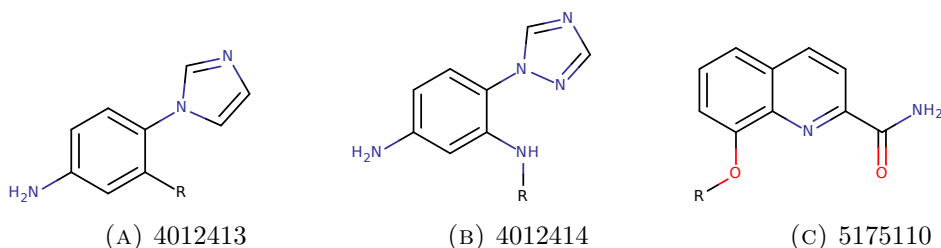


FIGURE 6.9: Illustration of the derived products from the three fragments used in our growing strategy.

pre-compiled growing). Nine products were suggested for synthesis via the *Mitsunobu phenol* reaction. The synthesis of the derived products is in progress.

In parallel, a structure-based strategy was applied by docking the S sample of SCUBIDOO into the PIM1 binding site. Several *low-hanging fruits* were identified and deconstructed, leading to the identification of building block 4012414 which made compelling interactions with the receptor. A quick similarity search allowed us to identify building block 4012413, a close analog of 4012414, but with optimized PSA. Dozens of derivatives were suggested for these two building blocks, and derivatives of 4012413 were scheduled for synthesis first, because 4012414 derivatives were considered more risky (i.e. all suggested products rely on an internal hydrogen bond, which might be unfavorable).

Interestingly, our strategy led to the identification of an active fragment (4012413), through an *in silico* screening. It was identified by exploiting the *bipartite product* philosophy. In this case, a *low-hanging fruit* was identified with one 'half' of the product making compelling interactions (fragment A), while the remaining part (fragment B) made suboptimal interactions. Derived products of the fragment A were then extracted from SCUBIDOO and docked into PIM1 in order to suggest molecules with an optimized extension (fragment B) for further synthesis. The synthesis of the derived compounds is in progress. However, 4012413, the base fragment itself was synthesized, tested using a thermal shift assay (shift of $+1.8^\circ\text{C}$), which could be taken as a first experimental hint that this fragment could bind to PIM1. The first attempt to crystallize this complex was successful, which could be taken as a second experimental hint. Crystals will be solved in the near future in order to hopefully validate the predicted binding mode of this fragment and give more insights for further optimization of this novel scaffold (i.e. the closest analog in PIM1 binders has a Tanimoto distance of 0.22 using ECFP4 fingerprints).

Even though synthesis results are in progress, this study already offers a *silver lining*. A fragment hit was identified using docking, as part of a bigger 'non-optimized' molecule. This can be seen as a camouflage, where we actually docked an efficient fragment disguised as a drug-like molecule. This could bring an alternative to the fragment docking

field, which remains challenging at many levels. One challenge is the number of poses that can be generated, due to the fact that fragments are small chemical entities that can bind at multiple different spots in multiple different fashions. On the top of that, possible interactions with water have to be taken into account and current docking program tools often do it implicitly (i.e. it is still one of the main challenge of docking tools). One could significantly reduce this number when disguising a fragment as a drug-like molecule, since one will add steric or polar constraints to the fragments. Such constraints will reduce the number of possible conformations that are energetically favorable when docked into a receptor, and thus reduce the number of generated poses.

In a drug discovery effort, this can be helpful. Indeed, a fragment will never be used as a drug by itself, simply because it is quite likely to be unspecific and exhibits only low affinity. Fragments are thus extended into bigger molecules to solve those problems and this process introduces new steric and/or polar constraints. Thus, docking SCUBIDOO products is somehow equivalent to docking fragments with additional steric and/or polar constraints that were assigned more or less randomly. The more derived products, the more likely we map the fragment *constrained space* delimited by the receptor. Thus, if derived products maintain consistent binding modes for their fragment *A*, one can take that as a hint that the fragment *A* is likely to have a favorable LE.

This notion is quite an important discovery and will need further examples in order to be validated. We started to do so in the following chapter, where I will describe the first large scale application of SCUBIDOO.

"So many vows... they make you swear and swear. Defend the king. Obey the king. Keep his secrets. Do his bidding. Your life for his. But obey your father. Love your sister. Protect the innocent. Defend the weak. Respect the gods. Obey the laws. It's too much. No matter what you do, you're forsaking one vow or the other."

Ser Jaime Lannister, *A Clash of Kings* (1998)

The chapter 7 is a scaffold of an article. I was responsible of the overall strategy, docking screening, visual inspection, selection of the fragments to optimize, purchasing the building blocks. Taros (Anna Karawajczyk) handled all the synthesis part.

Chapter 7

Tailored combinatorial synthesis guided by combinatorial growing

7.1 Introduction

Computational approaches play a crucial role in *de novo* design, helping to generate millions or even billions of virtual compounds within a short time frame. The enumerated *in silico* compounds allow computational or medicinal chemists alike to explore new sectors of the chemical space. While the novelty of the generated compounds can easily be assessed, the synthetic feasibility is often neglected [261], thus creating a hurdle that can be a barrier to further investigation. In order to circumvent this problem, we created SCUBIDOO (chapter 5), a database containing 21M virtual products which were optimized towards high likelihood of synthesizability. These virtual products are the assembly of two commercially available building blocks and were created using a collection of 58 robust organic reactions [1, 2]. SCUBIDOO aims at exploring new quadrants of the chemical space, which contain molecules that are, in principle, easier to synthesize than molecules with no predicted synthetic routes.

However, the synthetic feasibility of SCUBIDOO virtual products still needs to be validated in order to justify this database as a useful tool for ligand discovery efforts. With the recent publications of the “synthesis machine” [282], it will also be interesting to demonstrate that SCUBIDOO could be coupled with such automated robotic synthesis, in order to synthesize novel products at a large scale (i.e. more than 100 products). The challenge will then be to avoid usual combinatorial chemistry, but do tailored combinatorial synthesis that provide a diverse set of molecules that will enhance the chance to identify novel ligands for a given target.

In this study, we adapted the previous strategy where we designed potentially novel ligands for the PIM1 kinase (chapter 6) at a larger scale. We created a more elaborated “divide and conquer” strategy, where 240 virtual products were suggested for synthesis as likely novel ligands for the β_2 AR. In a first step (i.e. *division*), the binding site was divided in three regions: the orthosteric site, the secondary binding pocket (SBP) and the tertiary binding pocket (TBP). Two strategies were then defined aiming at growing a fragment from the orthosteric site towards the SBP (strategy 1) or towards the TBP (strategy 2). The second step (i.e. *conquest of the orthosteric site*) consisted of identifying 10 *A* fragments (5 for each strategy) which make compelling interactions with the orthosteric site, and find organic reactions compatible with each strategy). In the last step (i.e. *conquest of the SBP and TBP*), 48 *B* fragments (24 for each strategy) were selected so they are compatible with the defined organic reactions (i.e. compatible with growing) and offer divers binding modes.

This study strives to illustrate how SCUBIDOO can assist tailored combinatorial synthesis with the support of docking and combinatorial growing. 10 *A* fragments (5 + 5) were identified for further growing and were combined with 48 *B* fragments (24 + 24) using two different organic reactions, yielding 240 products (2 x (5 x 24)). The diversity of the *A* and *B* fragments was assisted by a polar fingerprint of the binding site (defined by listing all the polar residues present). The selection of the fragments and resulting products was guided by docking. The two organic reactions (*amide* formation and *reductive amination*) were selected due to their compatibility with the fragment growing strategies, but also because they represents half of the virtual products within SCUBIDOO. 127 products were successfully synthesized (53%), with an high synthesis success rate (82%) for the *amide* pool and a lower one (20%) for the *reductive amination* pool. Those preliminary results suggest that products from SCUBIDOO are amenable to automated robotic synthesis, and that the knowledge gathered at each iterative synthesis cycle could improve the overall synthesis success rate in the long run.

7.2 Methods

7.2.1 The beta-2 AR binding site

The β_2 AR binding site was previously described in the chapter 4 (PINGUI). In this study, we extended the binding site description towards three smaller cavities (figure 7.1): the orthosteric site, the secondary binding pocket (SBP) and the tertiary binding pocket (TBP).

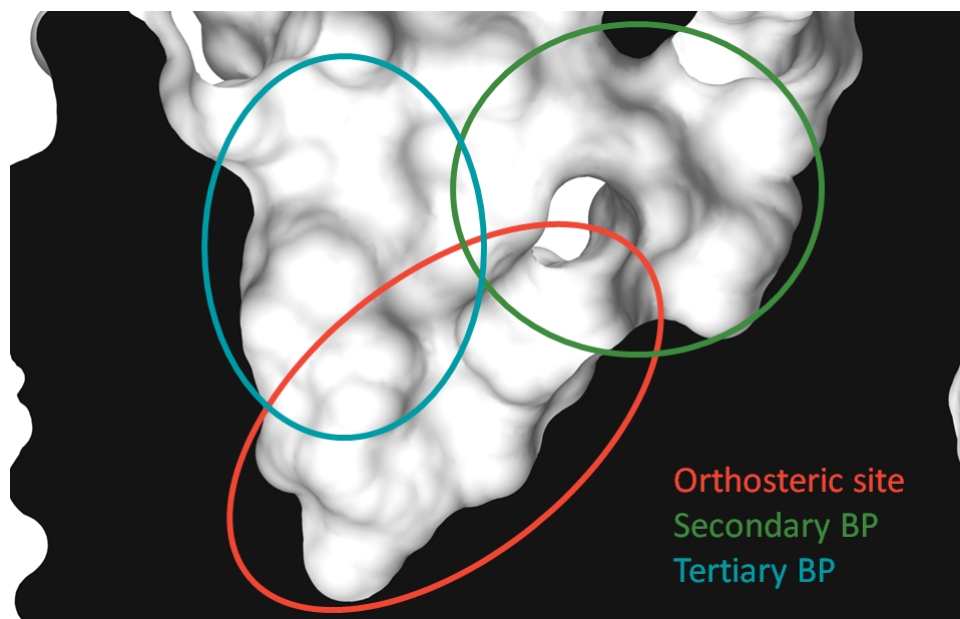


FIGURE 7.1: Sliced surface view of the binding site of the receptor in its active conformation (PDB code 2RH1). The orthosteric site is colored in red, the SBP in green and the TBP in blue.

7.2.2 Receptors preparation

Docking calculations were performed with the inactive conformation of the β_2 AR in complex with carazolol (PDB: 2RH1) [220, 221] and the active conformation in complex with a high affinity agonist (PDB: 4LDL, ligand = XQC) [226]. All ligands, solvent, lipid molecules as well as the T4-lysozyme insertion or the stabilizing nanobody Nb6B9 were removed. The hydrogens were placed and minimized using the HBUILD module in CHARMM [249].

7.2.3 Libraries preparation

The S sample and all the derivative libraries mentioned in this study were prepared for docking according to the same protocol:

- Compounds were protonated at physiological pH ($\text{pH} = 7.4$) and the most likely tautomer was assigned using QUACPAC [246].
- If any protonated tertiary amines were present, both enantiomers were generated using flipper [254].
- Up to 500 conformers were generated for each molecule by means of OMEGA [57].

7.2.4 Docking

All products from the samples and the derivative libraries were docked with FRED and HYBRID [58, 80–82] which is described in chapter 2 and 7. SZYBKI [87] was used as a refinement approach in order to minimize the poses found and remove clashes (internal or external).

7.2.5 Chemical descriptors

Calculator Plugins from ChemAxon [283, 284] were used for logD prediction and calculation. Chemical similarity was based on the Tanimoto score of the FCFP4 fingerprints [30], which were computed using the RDKit library [257]. Chemical descriptors used for the PCA analysis were computed using the RDKit library.

7.2.6 Datasets

The β_2 AR antagonist (204) and agonist (206) sets were downloaded from the GPCR Ligand Library (GLL) [285]. The drug-like subset was downloaded from ZINC [253] and contained 17'900'742 compounds. The *ether* pool comes from SCUBIDOO and was suggested as a third (backup) pool for synthesis.

7.2.7 Principal Component Analysis (PCA)

The *ether*, *reductive amination* and *amide* pools were compared to the antagonist and agonists sets using the PCA function as implemented in the R statistics environment. The descriptors used were molecular weight, logP, number of H-bond donors, number of H-bond acceptors, number of rotatable bonds, number of chiral center, topological polar surface area, fraction of aromatic ring and the Bertz index [286], which estimates the molecular complexity.

7.2.8 Synthesis

The 24 position Mettler-Toledo Miniblock contains 1 plate with 24 wells, each well is filled with one B building block. 1 plate is attributed for each of the 5 A fragment to grow, thus yielding 120 products (figure 7.2). This procedure is repeated twice, one for *reductive amination* and one for *amide* formation. The total procedure should yield up to 240 products (without stereoisomers taken into account).

7.2.8.1 1.5 g scale Boc-protection

Equipment required and reaction conditions: 3 necked round bottomed flask, magnetic stirrer and dry reaction under nitrogen.

Procedure: In a three necked round bottomed flask was placed the appropriate aminoacid (1.5 g, 1.0 eq) and diluted with 20 mL of MeOH. To the reaction mixture was added triethylamine (1.1 eq) and Boc-anhydride (2.0 eq) dissolved in 5 mL of MeOH was added dropwise over 5 minutes. The reaction was stirred at room temperature for 2-3 hours (starting material consumption was monitored by TLC).

Work up: The reaction mixture was evaporated to dryness. Crude was redissolved in EtOAc and washed twice with NaHCO₃. The aqueous layer was acidified with 10% HCl until pH = 2 and extracted three times with EtOAc. The combined organic layers were dried over MgSO₄, filtered and concentrated in vacuo to yield the corresponding Boc-protected aminoacids with moderate to excellent yields (38% - 92%).

7.2.8.2 40 mg scale amidification products

Equipment required and reaction condition: 24 position Mettler-Toledo Miniblock, magnetic stirrer and dry reaction under nitrogen.

Procedure: In a Mettler Vial were placed a previously prepared solution containing 40 mg of the corresponding Boc-protected aminoacid (1.0 eq) in 2 mL of DMF, DIPEA (5.0 eq), HOBT (1.5 eq), EDC*HCl (2.0 eq). Then the corresponding amine was added (1.0 eq). The reaction mixtures were stirred at room temperature overnight. Reaction conversion was confirmed through UHPLC check of some representative samples.

Work up: The reaction mixture was evaporated to dryness. Crude product was purified by preparative HPLC (gradient, Acetonitrile: water with 0.1% Formic acid, 2-98%). Fractions containing pure product were combined and evaporated to dryness in Mettler Vials.

7.2.8.3 De-Boc

Equipment required and reaction condition: 24 position Mettler-Toledo Miniblock, magnetic stirrer and dry reaction under nitrogen.

Procedure: Into a Mettler Vial containing the Boc-protected amidification product was added 0.5 mL of 1,4-dioxane and 0.5 mL of 4N HCl in dioxane. The mixtures were stirred at room temperature overnight. Reaction conversion was confirmed through UHPLC check of some representative samples.

Work up: The reaction mixture was evaporated to dryness. Crude product was purified by preparative HPLC (gradient, Acetonitrile: water with 0.1% Formic acid, 2-98%). Fractions containing pure product were analysed by UHPLC and some of them by ¹HNMR in selected cases.

7.2.8.4 30 mg scale reductive amination products

Equipment required and reaction condition: 24 position Mettler-Toledo Miniblock, magnetic stirrer and dry reaction under nitrogen.

Procedure: In a Mettler Vial was placed the appropriate amine (30 mg, 1.0 eq) and diluted with 2 mL of dry DCE. To this solution was added appropriate aldehyde (0.9 eq) and acetic acid (1.5 eq). The reaction was stirred at room temperature for 20 minutes and then sodiumtriacetoxyborohydride (1.5 eq) was added. The mixtures were stirred at room temperature overnight. Reaction conversion was confirmed through UHPLC check of some representative samples.

Work up: The reaction mixture was washed with 1 mL of water and the organic layer was evaporated to dryness. Crude product was purified by preparative HPLC (gradient, Acetonitrile: water with 0.1% Formic acid, 2-98%). Fractions containing pure product were analyzed by UHPLC and some of them by ¹HNMR in selected cases.

7.2.9 Selection strategy

Several criteria had to be taken into account for the 240 compounds selection. Each criterion will be illustrated and more thoroughly explained in the Results section. The list below was the pillar that defined our “divide and conquer” strategy and should be regarded as a general guideline.

- Which part of the binding site should be aimed for ? Which residues ?
- Which reactions should be applied ?

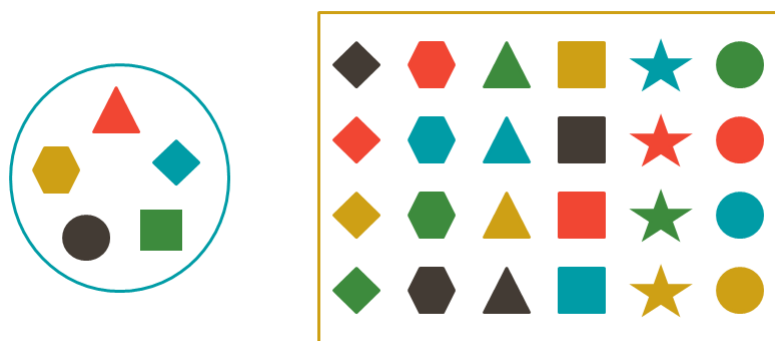


FIGURE 7.2: Parallel synthesis illustration. One plate containing 24 B building blocks (within the yellow rectangle) is reacted against one A fragment yielding 24 products. This procedure is repeated five times, one for each A fragment to grow (within the blue circle).

- Which 5 A building blocks ?
- Which 24 B building blocks ?
- Safeness, boldness and recklessness of the suggestions
- Chemical diversity and novelty
- Price and availability

7.3 Results

7.3.1 Binding site partitioning

As mentioned earlier, the binding site was divided in three sub-cavities: the orthosteric site, the SBP and the TBP (figure 7.1). This partitioning fits nicely with the SCU-BIDOO philosophy which is based on bi-partite products, and one can imagine each sub-cavity being filled with one building block. Thus, we decided to applied two different strategies: one aiming at filling the orthosteric site and the SBP (a follow up application of PINGUI), and the second one aiming at filling the orthosteric site and the TBP.

7.3.2 Polar residue fingerprint

All polar residues of the binding site were exhaustively enumerated in order to create a *polar map* of the possible interactions (figure 7.3). This map can be seen as a simple fingerprint encoding the polar residues in the binding site (figure 7.4). Hydrophobic

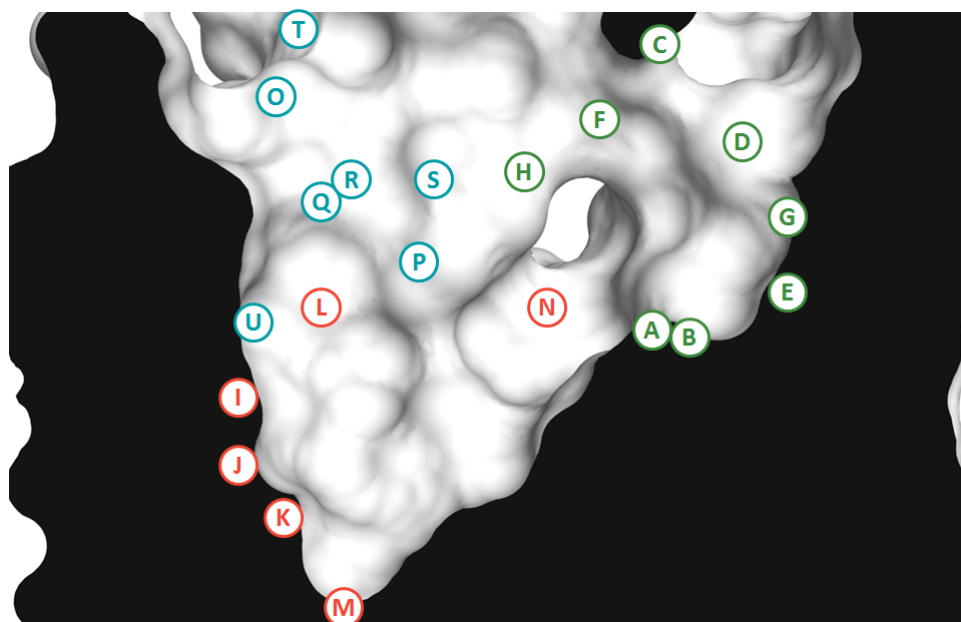


FIGURE 7.3: 2D depiction of the beta-2 AR binding site with all polar residues.



FIGURE 7.4: Polar fingerprint of the beta-2 AR binding site.

residues were omitted because they are generally involved in non-specific interactions [3] and we wanted to focus mainly on specific (i.e. polar) interactions. During the selections of the products, this fingerprint was utilized to suggest divers compounds (i.e. we tried to interact with as many polar residues as possible).

7.3.3 Identification and selection of the A fragments to grow

A typical workflow for the use of SCUBIDOO in structure-based efforts is illustrated in figure 6.5 in chapter 6. Each step will be detailed below.

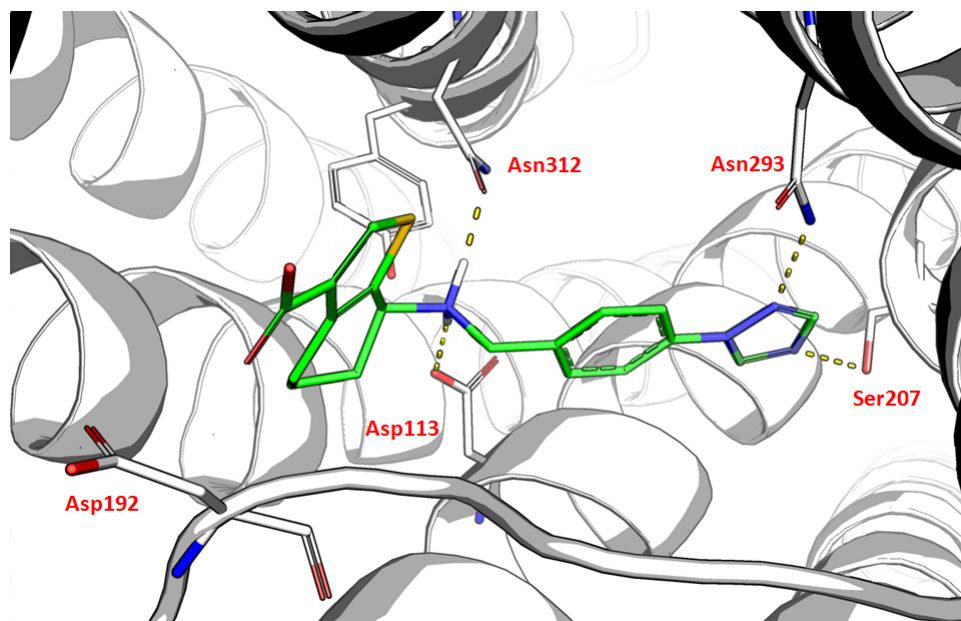


FIGURE 7.5: Low-hanging fruit identified in the virtual screening of the beta-2 AR with the SCUBIDOO S sample.

7.3.3.1 Step 1: sample docking.

The S sample was docked in the binding site of the receptor in its active and inactive conformation using FRED. For each docking run, the top 500 molecules based on the *score* were visually inspected, in order to identify *low-hanging fruits* (i.e. products containing building blocks that make optimized interactions with the protein). Compound 18806850 is a typical *low-hanging fruit* (figure 7.5) containing both an optimized and unoptimized building block. The optimized building block contains a protonated amine moiety interacting with the crucial Asp113^{3,32} and it also contains a triazole moiety which engages in favorable interactions with Asn293^{6,55} and Ser207^{5,46}. The unoptimized building block contains an acid moiety which is not involved in any interactions, thus penalizing the resulting compound.

7.3.3.2 Step 2: deconstruction of the low-hanging fruits.

Each *low-hanging fruit* was deconstructed in order to identify the reaction and building blocks necessary for its predicted synthesis. This procedure allowed to identify building blocks compatible with *reductive amination* or *amide* formation that could be starting points (Fragments A) in our combinatorial growing. For instance, compound 18806850 was predicted to be synthesized using building blocks 5130031 and 4012456 via *reductive amination*. Building block 4012456 was selected for the next step, due to its most favorable interactions with the orthosteric site.

7.3.3.3 Step 3: selection of the organic reactions

Two organic reactions were selected, one for each growing strategy. *Reductive amination* was applied to fill the orthosteric site and the SBP, because it showed positive results in PINGUI. *Amide* formation was used to explore the orthosteric site and the TBP.

7.3.3.4 Step 4: construction of the building block derivatives.

The derived products of each promising building block based on *reductive amination* or *amide* formation were downloaded from SCUBIDOO.

7.3.3.5 Step 5: docking derivatives.

The derived products were then docked into the binding site of the receptor in its active and inactive conformational state. Five *A* fragments were selected for each reaction, yielding ten *A* fragments for the combinatorial growing. The selection of the *A* fragments was based on the propensity of matching a consistent binding mode among the different derivatives. This procedure was done visually by inspecting the top 500 molecules based on the *score* and repeated for each *A* fragment.

7.3.3.6 Step 6: mini SAR around the A fragments.

For each reaction, in order to facilitate the next step (i.e. selection of the 24 *B*), the five *A* fragments were also selected based on their binding mode (i.e. as dissimilar as possible) and similarity (i.e. two pharmacophores at the same distance) as illustrated in figure 7.7 and 7.6.

For the *reductive amination* derivatives, the pharmacophores were defined by an amine and an aromatic moiety separated by two carbons (i.e. epinephrine-like). Such pharmacophore filters are important to ensure the amine to be located in the same position. Thus extensions that will be attached (24 *B*), should also be located in the same site. Only compound *E* in figure 6.7 differs within the pool, but the position of the amine was predicted to be in the same vicinity as the other fragments (figure 7.8). Compounds 1 to 3 represent a range of variations of the pyridine / pyrimidine scaffolds, with a simple methyl (compound 1) mostly doing hydrophobic contacts, a hydroxy moiety (compound 2) predicted to interact with Ser207^{5.46} and a hydroxy with an aromatic amine moiety (compound 3) predicted to interact with both Ser207^{5.46} and Ser203^{5.42}.

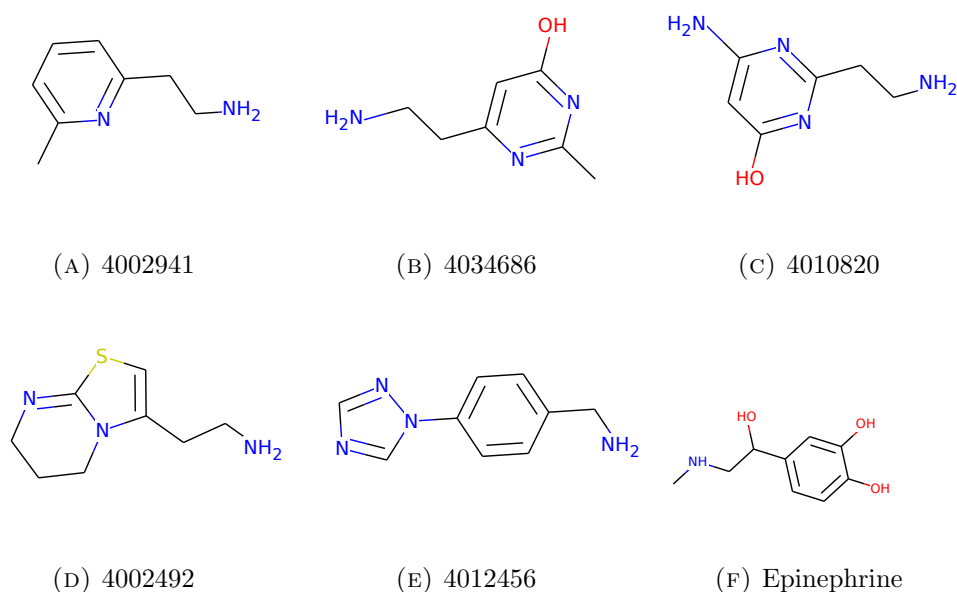


FIGURE 7.6: (a-e) Five fragments were selected for combinatorial growing in the *reductive-amination* pool. (f) Epinephrine.

For the *amide* pool, the pharmacophore aimed to reproduce the beta-hydroxy motif but replacing it with a ketone group, thus losing the donor property. Among the 5 *A* fragments, subtle difference were taken into account for the selection of divers yet close molecules:

- Variation of the distance between the amine and the ketone moieties: two carbons for compounds 4140254 and 4027710 and three carbons for the remaining ones.
- Variation on the cycle size (from 5 to 7 atoms).
- One primary amine (4045566) and four secondary amines were selected.

7.3.4 Identification and selection of the B building blocks to attach

For a given reaction pool, the 24 *B* building blocks were selected in order to create a cocktail of *safe*, *bold* and *reckless* chemical features.

The *safe* building blocks were selected by applying the concept of *molecular obesity* [3] in a *reverse-engineering* strategy. Molecular obesity works by adding some *grease* (i.e hydrophobic moieties) to a potent molecule in order to improve its potency via hydrophobic interactions and a higher MW which will decrease the desolvation penalty. In the context of this study, the main goal was to identify fragments with high ligand efficiency and then optimize them into high affinity ligands. Thus, in order to increase

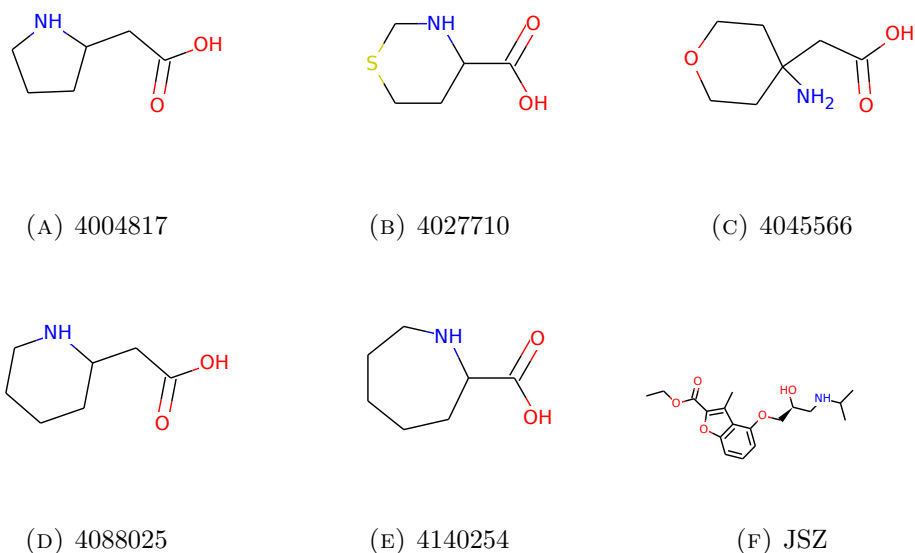


FIGURE 7.7: (a-e) Five fragments were selected for combinatorial growing in the *amide* pool. (f) JSZ, an high affinity inverse agonist.

the chance of identifying efficient fragments, one needs to attach some *grease* to them so they are more likely to show up in the assays. *Safe* extensions were thus defined when no polar atoms were present, as illustrated in figure 7.8.

Bold building blocks were defined as those containing one polar atom predicted to be involved in a polar interaction. In a similar fashion, *reckless* building blocks contain two or more polar atoms. The assumption being that polar interactions are hard to form because they are highly directional. Thus the more polar interactions need to be satisfied, the more likely the compound is to fail as active ligand (i.e. each unsatisfied polar atom induces desolvation penalty, free H-bond donors induce a higher penalty). Building blocks were selected so they map to as many polar interactions as defined by the binding site fingerprint. Doing so allows one to explore which residues might be implicated (or not) in the activation of the receptor as well as ensuring chemical diversity among the suggested molecules.

7.3.4.1 Amination

The 24 *B* building blocks of the *reductive amination* pool are illustrated in SI 7.5. For this pool, it was difficult to find *safe* extensions. Thus we made use of the in house library of Taros and selected 8 buildings blocks that were attached to the *A* fragments using PINGUI (chapter 4). Among the 24 suggested building blocks, 7 were selected as *safe* extensions (ID = 4004292, 4004691, 5100225, TCR00002575, TCR00002580, TCR00002587 and TCR00002670), 15 as *bold* and 2 as *reckless* (ID = 4002557, 4301989).

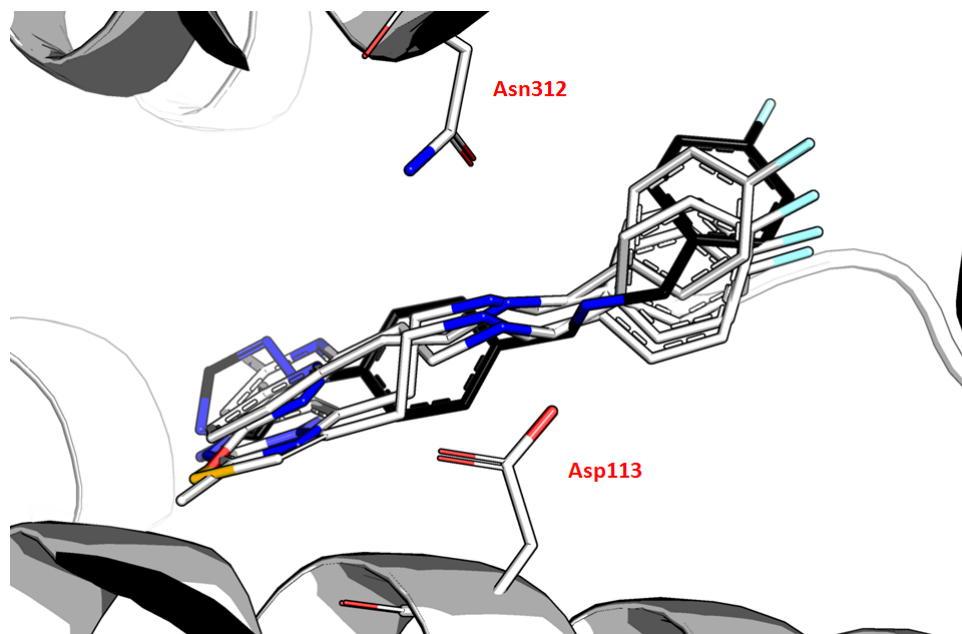


FIGURE 7.8: Illustration of molecular obesity with a *safe* (hydrophobic) building block which keeps a consistent binding mode among 4 A fragments (white carbons).

An example of a *safe B* building block is illustrated in figure 7.8, where the chlorobenzene keeps a consistent binding mode with 4 of the 5 A fragments (i.e. the ones that share a common pharmacophore).

7.3.4.2 Amide formation

The 24 *B* building blocks of the *amide* pool are illustrated in SI 7.5. 5 were selected as *safe* (ID = 4002741, 4002969, 4006408, 4028089, 4028827), 14 were selected as *bold* and 5 were selected as *reckless* (ID = 4040087, 4031726 (internal H-bond), 4037223, 4045370, and 4045452).

7.3.5 Analysis of the generated products

7.3.5.1 Principal Component Analysis

Analysis of the space spanned by the aforementioned physicochemical properties (Figure 7.9) using the two main principal components depicts the suggested products as overlapping with the property regions of known agonists and antagonists of the β_2 AR. Interestingly, the *amide* pool overlaps with a region of known antagonists where no agonists are present. This could suggest that the *amide* pool contains products with an antagonistic profile. This is supported by the fact that the *amide* pool explores the TBP which is not present in the active conformation of the β_2 AR. The *reductive amination*

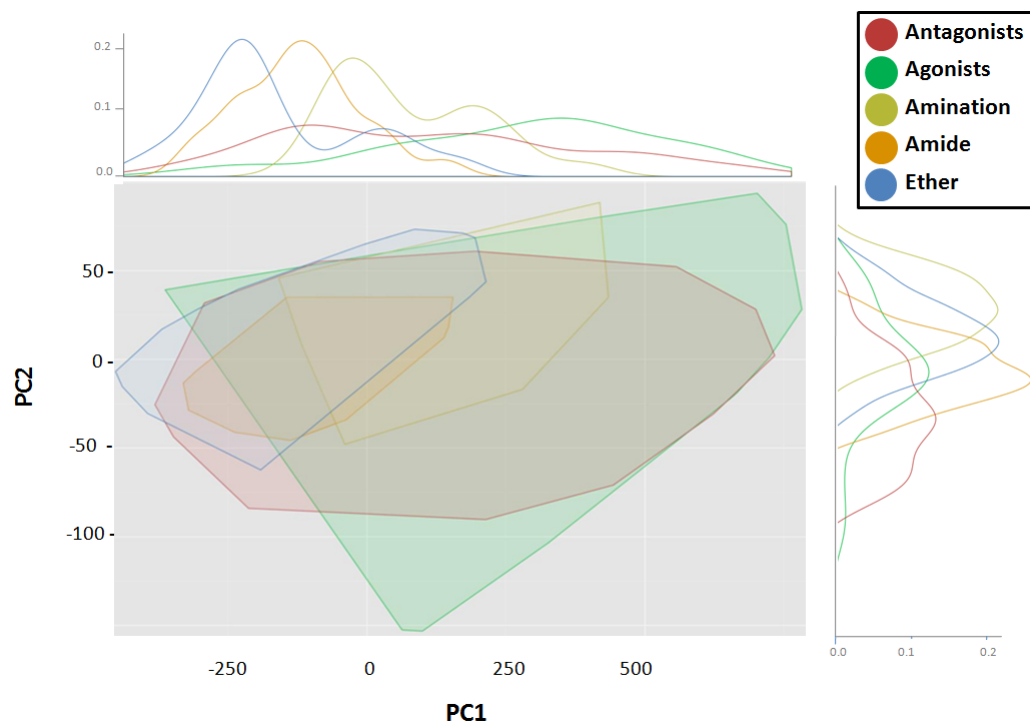


FIGURE 7.9: PCA of the amide pool (orange), amination pool (light green), ether pool (blue), 206 known agonists (green) and 204 known antagonists (red) of the β_2 AR.

pool offers similar behavior and overlaps with a region of known agonists where no antagonists are present. This could suggest that the *reductive amination* pool contains products with an agonistic profile. This is supported by the fact that many agonists from the β_2 AR interact with the SBP, highlighting the likely role of this pocket in the activation of the receptor.

7.3.5.2 Novelty

Each product of the *amide* and the *reductive amination* pools were compared with the GLL (206 agonists and 204 antagonists) and the ZINC drug-like set in order to retrieve the closest analog based on the ECFP4 fingerprints. Both pools are dissimilar to known actives (Tanimoto distribution centered around 0.17 and 0.21 for the *amide* and the *reductive amination* pools respectively) suggesting that the suggested compounds contain novel scaffolds that could provide new insights in the β_2 AR activation mechanism. When compared to known molecules (i.e. drug-like set) no matches with a Tanimoto similarity of 1 were retrieved, highlighting that all the suggested compounds are truly novel chemical entities.

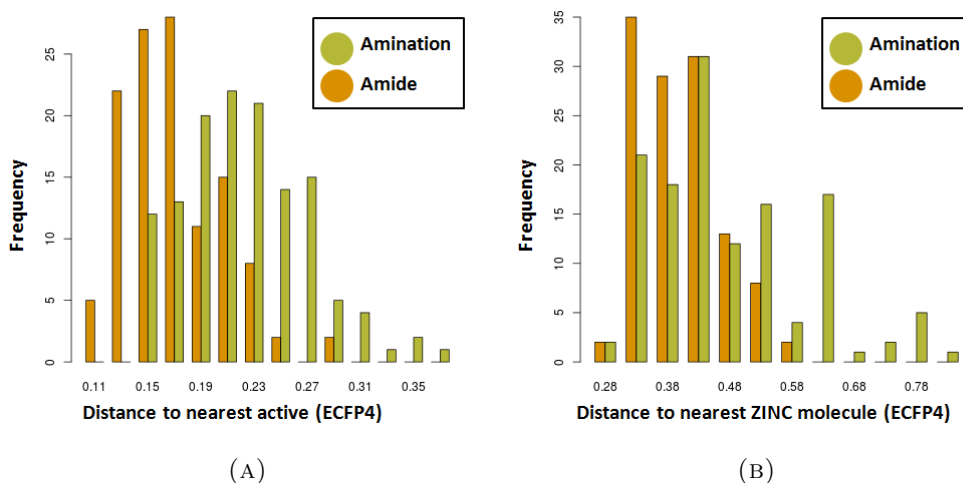


FIGURE 7.10: ECFP4 Tanimoto score distributions of the *amide* and the *reductive amination* pools compared with (A) the GLL and (B) the ZINC drug-like set.

7.3.6 Synthesis

After completion of the allocated synthesis, purification and structural characterization activities for the project, a total of 127 final compounds have been delivered (53% synthesis success). 102 are part of the *amide* pool (82% synthesis success), while 25 belongs to the *amination* pool (20% synthesis success). The purity and amount of the synthesized compounds is gathered in figure 7.11.

The high synthesis rate (82%) of the amide pool is quite remarkable considering that this was a 3-steps reaction (boc - coupling - deboc). Interestingly, all five attempts to grow building block 4028089 failed, while only one attempt to grow building block 4033241 (i.e. a close analog of 4028089) failed. This result is quite odd given how close the two building blocks are (MACCS Tanimoto = 0.94) and thus one could have expect similar synthetic yield.

On the other hand, the synthesis rate of the amination serie was quite low (20%). This could be explained by the fact, that all the fragments to grow (i.e. the five *A* fragments) are primary amine. During the *reductive amination*, the aldehyde (i.e. any of the 24 *B* fragments) was added slowly so it can react and form the final product, which is a secondary amine. However, secondary amines can still react under *reductive amination* conditions and form tertiary amines. In case were both primary and secondary amines are present, primary amines should be favored for coupling, because they have an higher basicity than secondary amines. The difficulty of separating secondary and

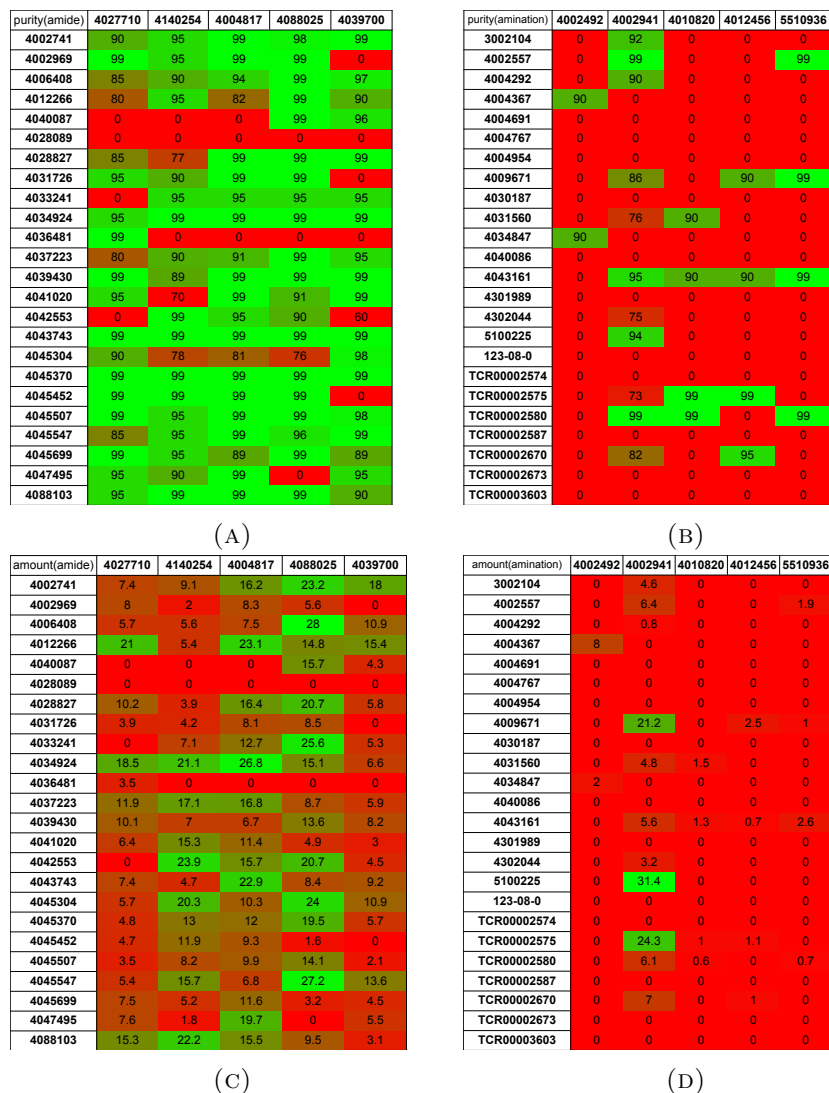


FIGURE 7.11: Heat map illustrating the purity and amount of the synthesized compounds by Taros. The purity of each product in the amide pool (A) and the amination pool (B) is expressed in % after LC-MS purification. The amount of each product in the amide pool (C) and the amination pool (D) is expressed in mg.

tertiary amines partially explains the low synthetic yield. Furthermore, most of the suggested products are highly polar hampering the purification process even more, thus yielding a few (25) purified products.

7.4 Discussion

We have introduced the second application of SCUBIDOO in a ligand discovery effort focusing on the β_2 AR. This study is in the continuity of chapter 6 but represents a leap forward in term of means, since SCUBIDOO was coupled with automated robotic synthesis which can handle two dozens of synthesis in parallel. Thus, this work enhance

the previous foundations describing how to use SCUBIDOO, by tackling the challenging task of suggesting products for tailored combinatorial chemistry guided by combinatorial growing and docking. The overall workflow can be assimilated as a “divide and conquer” strategy. The *division* consists in partitioning the binding site into smaller regions and map out all possible polar interactions. The *conquest* is made of two phases. First, identify *A* fragments with consistent binding mode that will be the core of a growing strategy and find organic reactions compatible with the envisioned growing. Second, identify extensions (*B* fragments) that can be combined with *A* fragments and which could yield to virtual products with favorable binding modes.

During the *division*, the binding site of the β_2 AR was partitioned in three regions: the orthosteric site, the SBP and the TBP. By making use of the bipartite philosophy (i.e. every product is the assembly of two building blocks), two different strategies were then defined. The first one strove to grow fragments from the orthosteric site towards the SBP, while the second strategy aimed to grow fragments from the orthosteric site towards the TBP. This *division* strategy ensures that each region is explored. In order to improve the exploration of these regions, a polar fingerprint was defined which list all the polar residues present in those regions. Doing so helps in suggesting chemically diverse building blocks, by ensuring that they interact with different residues. Ensuring such diversity could also help in identifying new residues involved (or not) in the activation of the β_2 AR.

The *conquest* of the orthosteric site was guided by a structure-based strategy, where the S sample of SCUBIDOO was docked in the β_2 AR binding site in its active and inactive conformations. 10 *A* fragments were identified with favorable binding mode in the orthosteric site and compatible for further growing. Among those 10 *A* fragments, 5 fragments were compatible with *reductive amination* and allowed one to grow the fragments towards the SBP, while the 5 remaining fragments were compatible with *amide* formation and thus growing towards the TBP was feasible. Chemical diversity was ensured within the *reductive amination* pool, so those fragments are likely to interact with Ser207^{5.46} and Ser203^{5.42} with different chemical interactions (charged amine, hydroxy, amine, aromatic nitrogen). The *amide* pool consists of close analogs with subtle changes (size of the ring varies from 5 to 7, distance between the amine and the acid moiety varies from 2 to 3 atoms) in order to increase the chance of the ketone group to interact with Asn312^{7.39}. Finally for each pool, a pharmacophore was defined to increase the chance of the reactive group (i.e. where the growing occurs) to be located in the same region. Doing so, allow the extensions attached to the core fragments to have more chance to have a consistent binding mode among each other, and thus increase our chance to identify bioactive compounds.

The conquest of the SBP or TBP was also guided by docking. For each *A* fragments to grow, the derived products based on the reaction of interest were docked in the β_2 AR binding site. For each pool (i.e. *reductive amination* or *amide* formation), a consensus of the best top 500 of the *B* fragments yielded the final 24 *B* fragments for tailored combinatorial synthesis. This selection was also guided by a classification of the *B* fragments into three categories: *safe*, *bold* and *reckless*. This classification relies on the number of H-bond interactions to satisfy. After the experimental assays, we aspire to show that the 'safe' series yield a higher hit rate than the 'bold' or 'reckless' series.

The automated robotic synthesis yielded 127 products out of 240 (53% synthesis success rate). 102 products came from the *amide* pool (82% synthesis success) while only 20 products came from the *reductive amination* pool (20% synthesis success). In both case, we learned a lot and already plan to implement new warnings within SCUBIDOO in order to facilitate next synthesis.

As instance, for the *amide* pool, the presence of an amine group and an acid moiety within each *A* fragment, make them likely to react with their self under *amide* formation conditions. Thus, our chemist collaborator suggested a Boc protection (and deprotection) in order to synthesize the products. We plan now to implement a new visual warning in the database for such building block reacting under those conditions. This will be coupled with a new tables suggesting potential protection groups and should help to improve the experience of future users.

Furthermore, with the *reductive amination* pool, we highlighted that our products were highly polar which complicated the purification step, yielding a low synthetic success rate (20%). Thus more rigorous filters (solubility or logD) could be applied earlier on, in order to increase the synthesis rate. It has to be noted that a real chemist would have probably obtain an higher synthesis rate if he would have tried to synthesize each product individually. This would have require several iterations for each compound, and probably different reaction conditions. This would have take probably months and several organic chemists to do so. In our scenario, the synthesis and purification was achieved within a few days, and only two iterations.

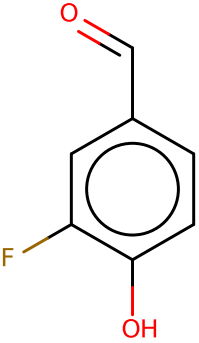
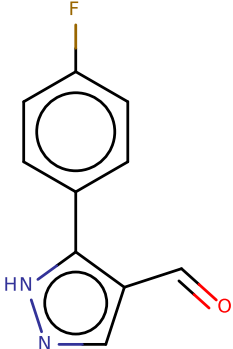
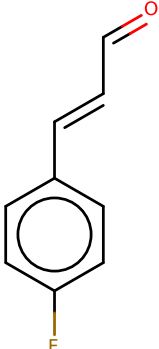
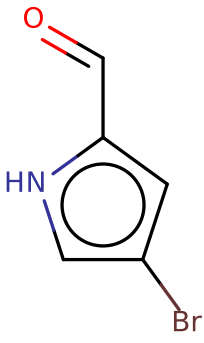
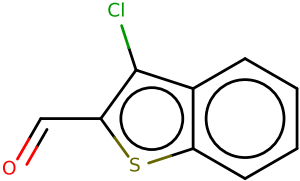
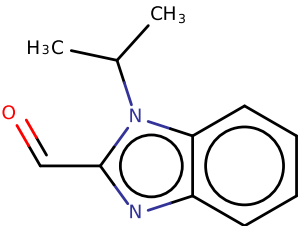
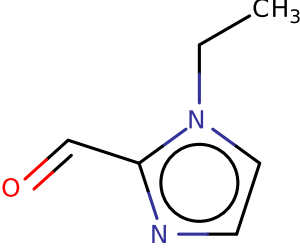
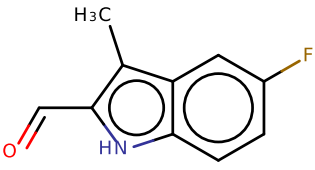
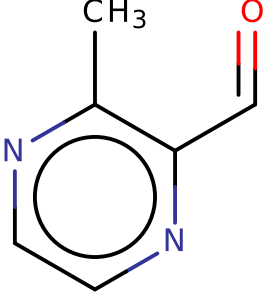
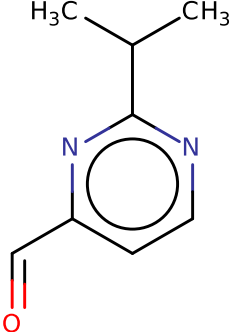
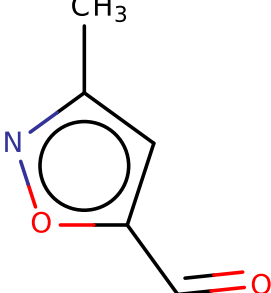
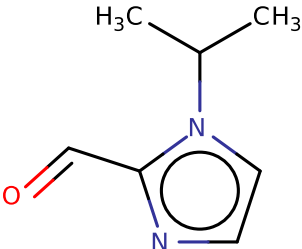
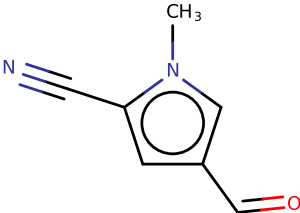
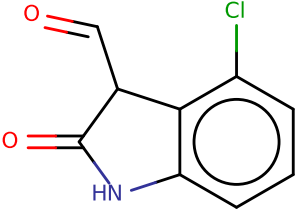
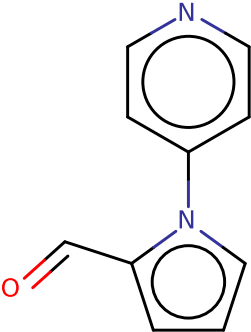
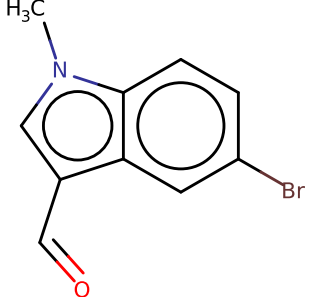
With a 53% synthesis success and a lot of lessons learned, this study suggests that SCUBIDOO is amenable to be integrated to automated robotic synthesis. Every synthesis attempt is prone to improve the knowledge contained within the database and thus increase the synthesis success rate over time. 50% of the virtual products within SCUBIDOO are made from either *amide* formation or *reductive amination*. One could thus argue that, even though 240 products were suggested for synthesis, these 240 virtual products could actually represent a larger fraction within the database.

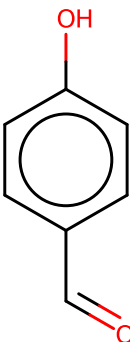
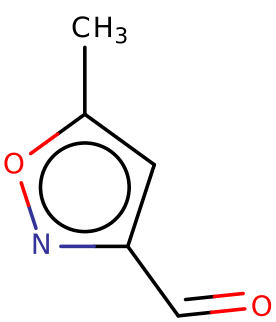
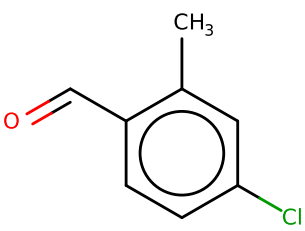
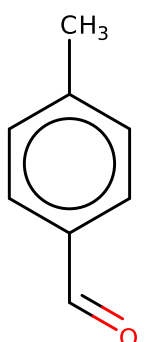
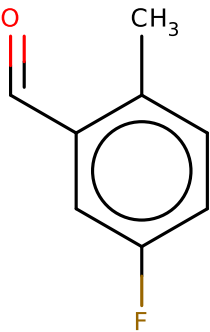
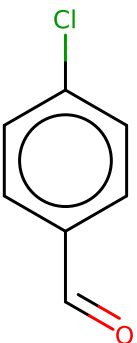
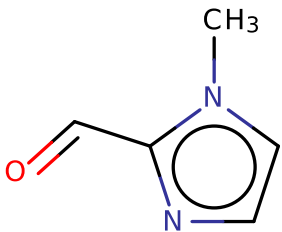
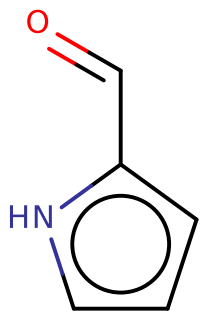
Chemical similarity analyses showed that all the suggested compounds were truly novel chemical entities, thus demonstrating how SCUBIDOO can explore new quadrants of the chemical space. Further comparisons to known actives of the β_2 AR, highlighted that the *amide* pool could contain molecules with an antagonistic profile, while the *reductive amination* pool could contain molecule with agonistic properties. This speculation still needs validation, that will come with further experimental assays (planned for September / October 2016).

The assays of the 127 products will help to hopefully identify novel ligands for the β_2 AR and validate our aforementioned assumption. Furthermore, we plan to make the dataset available (activities, synthetic yields, purity) since this pool could be a wonderful source of informations for QSAR or QSPR models. Indeed, all molecules are close to each other and such interconnected datasets could allow the scientific community to create and validate more accurate local models with high discriminative power.

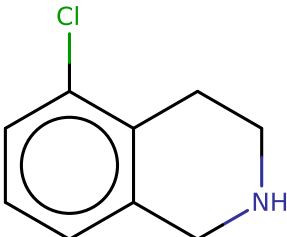
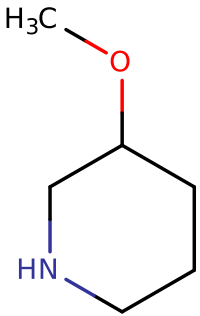
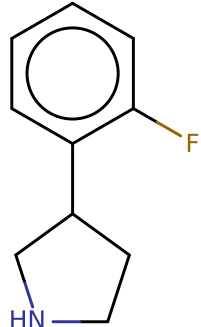
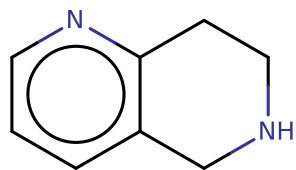
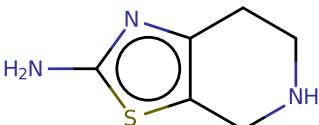
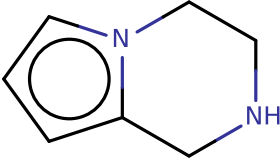
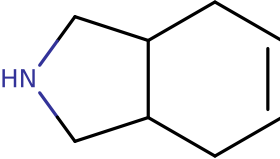
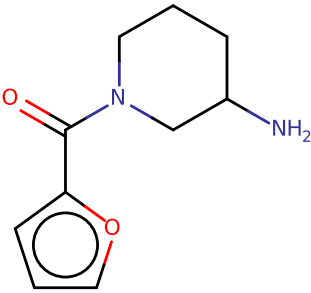
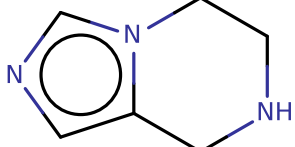
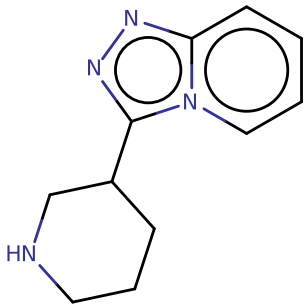
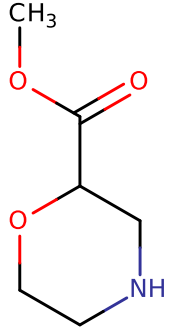
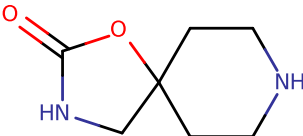
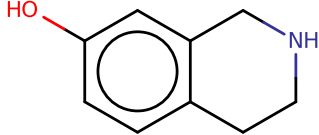
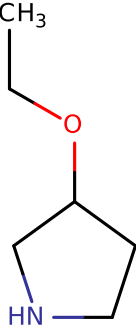
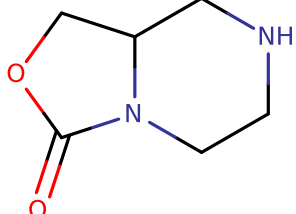
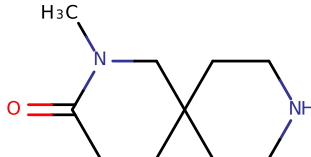
7.5 Supporting Information

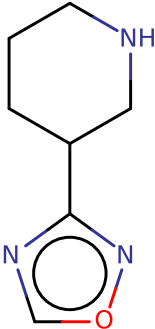
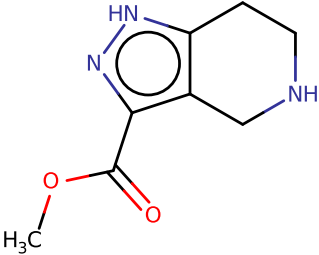
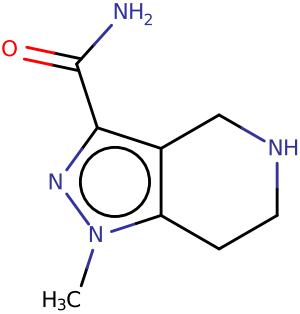
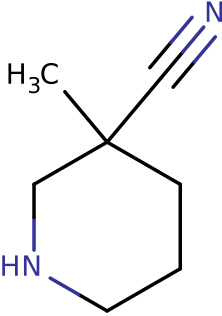
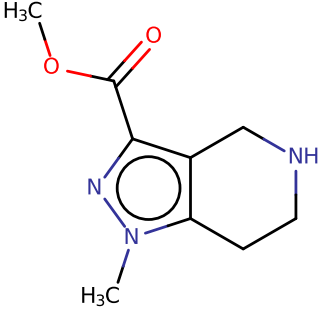
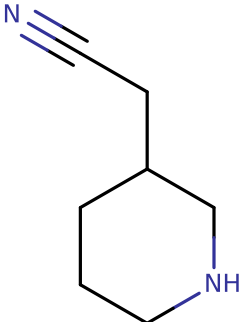
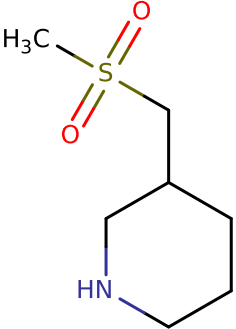
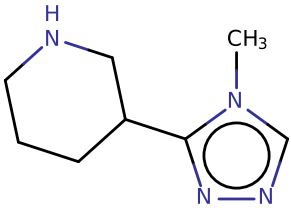
7.5.1 B building blocks for the reductive amination pool

<p>1</p>  <p>3002104</p>	<p>2</p>  <p>4002557</p>	<p>3</p>  <p>4004292</p>	<p>4</p>  <p>4004367</p>
<p>5</p>  <p>4004691</p>	<p>6</p>  <p>4004767</p>	<p>7</p>  <p>4004954</p>	<p>8</p>  <p>4009671</p>
<p>9</p>  <p>4030187</p>	<p>10</p>  <p>4031560</p>	<p>11</p>  <p>4034847</p>	<p>12</p>  <p>4040086</p>
<p>13</p>  <p>4043161</p>	<p>14</p>  <p>4301989</p>	<p>15</p>  <p>143 4302044</p>	<p>16</p>  <p>5100225</p>

<p>17</p>  <p>123-08-0</p>	<p>18</p>  <p>TCR00002574</p>	<p>19</p>  <p>TCR00002575</p>	<p>20</p>  <p>TCR00002580</p>
<p>21</p>  <p>TCR00002587</p>	<p>22</p>  <p>TCR00002670</p>	<p>23</p>  <p>TCR00002673</p>	<p>24</p>  <p>TCR00003603</p>

7.5.2 B building blocks for the amide pool

1	2	3	4
			
4002741	4002969	4006408	4012266
5	6	7	8
			
4040087	4028089	4028827	4031726
9	10	11	12
			
4033241	4034924	4036481	4037223
13	14	15	16
			
4039430	4041020	146 4042553	4043743

<p>17</p>  <p>4045304</p>	<p>18</p>  <p>4045370</p>	<p>19</p>  <p>4045452</p>	<p>20</p>  <p>4045507</p>
<p>21</p>  <p>4045547</p>	<p>22</p>  <p>4045699</p>	<p>23</p>  <p>4047495</p>	<p>24</p>  <p>4088103</p>

"The disease is where the drugs are not."

Fire in the blood (2013)

Chapter 8

Perspectives

8.1 Preamble

In the course of the previous chapters, we first developed an *in silico* workflow aiming at growing fragments towards synthetically feasible products (chapter 4). The synthesis success highlighted the applicability of the organic reactions set compiled by Hartenfeller et al. [1, 2]. We then decided to fully utilize this set and created SCUBIDOO, a database of virtual products generated by combining commercially available building blocks with the reaction set (chapter 5). This database concept was then applied in two different ligand discovery efforts, one focusing on PIM1 inhibitors (chapter 6) and the other on β_2 AR ligands (chapter 7). The initial goal of these two endeavors was to demonstrate the synthetic feasibility of SCUBIDOO products and the applicability of the database to fragment-based ligand discovery efforts. However, along the way we started to learn how to use SCUBIDOO and uncovered some unexpected strategies, which I think could be supportive for fragment-based ligand discovery efforts in general, and most notably for the field of fragment docking. Four noteworthy discoveries will be further detailed below:

- The bipartite product philosophy and its application to docking.
- Molecular obesity applied to fragment rescue.
- Tailored combinatorial synthesis guided by combinatorial growing and docking.
- Creation of the 'loser bracket' to rescue optimal chemical moieties identified through docking.

I will finish with a last paragraph on SCUBIDOO's future, in order to discuss the perspectives.

8.2 Bipartite products and docking disguised fragments

Docking of fragments remains challenging. As we have learned from experimental approaches, fragments can bind to multiple sites in a receptor. Computational approaches reflect that, by usually providing diverse solutions for a given fragment. Program tools for docking are reliable to predict the correct binding modes. However, their weakness is the scoring phase (and thus ranking) and the fact that they do not take into account water molecules which play a crucial role upon binding (e.g. mediated water interactions). This makes the process of identifying correct binding modes harder, as we discussed in chapter 2.

As illustrated in chapters 5, 6 and 7, SCUBIDOO could help to suggest more reliable docking poses for fragments, by means of the *bipartite product philosophy*: “any product is the assembly of two building blocks”. Thus, docking products from SCUBIDOO can be considered equivalent to docking one fragment disguised as a drug like compound (i.e. a sort of camouflage). In other words, even a docked product with some unfavorable interactions (i.e. a *rotten fruit*) could actually hide an optimal fragment (i.e. a *tree*). Indeed, if half of the product shows compelling interactions (i.e. high LE), the second half doesn’t really matter and could be assimilated as steric or polar constraints that can be optimized later on (exploration of the *tree*: figure 6.5).

In a drug discovery context, this could be helpful. A fragment on its own will never be used as a drug, simply because it is likely to be unspecific and exhibit a low affinity. Fragments are therefore usually extended into larger molecules to solve these problems and this process introduces new steric and/or polar constraints. SCUBIDOO natively answers this problem, because all the virtual products are fragments extended with additional steric and/or polar moieties (i.e. already-grown fragments).

Furthermore, docking a fragment with additional constraints considerably reduces the number of enumerated poses. In an early fragment-based ligand discovery effort, this could help to prune all possible binding modes to the more plausible ones. For instance, if the predicted binding mode of a *A* fragment on its own is unreliable (i.e. several different suggestions), one can turn to derivatives of fragment and dock them. Then, if the derived products maintain consistent binding mode for the *A* fragment, one can take it as hint that the fragment is more likely to have a reliable binding mode. This scenario is illustrated in chapter 7, where the 10 *A* fragments for growing were selected based on the docking poses from their derivatives.

Docking bipartite derivatives makes the visual inspection more friendly and to a certain extent more reliable than standard drug-like molecules. When dealing with derivatives from the same fragment *A*, one generally expects the fragment *A* to bind to the same

region (i.e. same binding mode). If not, one could assume that the initial binding mode is not consistent, thus the fragment *A* binding mode is not trustworthy enough and one could quickly discard such a fragment. On the other hand, if it preserves a persistent binding mode, one could take that as a hint that it could bind efficiently at this position. In addition, it facilitates the visual inspection, because one only has to focus on half of the product.

8.3 Reverse engineering via molecular obesity

We learned in chapter 4 that suggesting hydrophobic extensions to our initial fragments was more likely to yield a leap in affinity. This is also a good illustration of *molecular obesity*. Indeed, we added a bit of 'grease' (i.e. hydrophobic moieties) to our fragments, and by doing so, we increased the molecular weight without increasing the PSA. From an entropic point of view, the resulting product will be easier to desolvate. Since it is known that desolvation plays an important role in binding [124], this aligns with the improved affinities for the products. The downside for this study was that we didn't learn that much about important residues that might be involved in the activation of the β_2 AR. Indeed, most of the extensions were making non-specific interactions. In summary, this study stressed that an 'easy' way to improve the affinity of a fragment is to extend it by means of hydrophobic moieties.

This strategy could also be applied retroactively, in a *reverse engineering* fashion. Thus, in order to increase the chance of identifying fragment hits, one could test fragments extended with hydrophobic moieties. In the context of bipartite products, this means that half of the product makes optimized interactions (*specific head*) while the remaining half makes likely non-specific interactions (*unspecific tail*). This could also help to rescue fragments from both *in vitro* and *in silico* screenings, and thus allow one to identify fragments that would not have been picked up if tested/screened on their own.

At the moment, rescuing fragments with the help of molecular obesity is purely speculative, but we have taken measures so we could hopefully validate this hypothesis. In chapter 7, 240 molecules were designed so that they bind to the β_2 AR. 120 molecules were synthesized by means of *reductive amination* and among those molecules, 35 'safe' molecules consist of one *specific head* and one *unspecific tail*. We aspire to show that the 'safe' series yield a higher hit rate than the 'bold' or 'reckless' series. The seven 'obese' building blocks which yielded the 35 'safe' molecules are illustrated in figure 8.1.

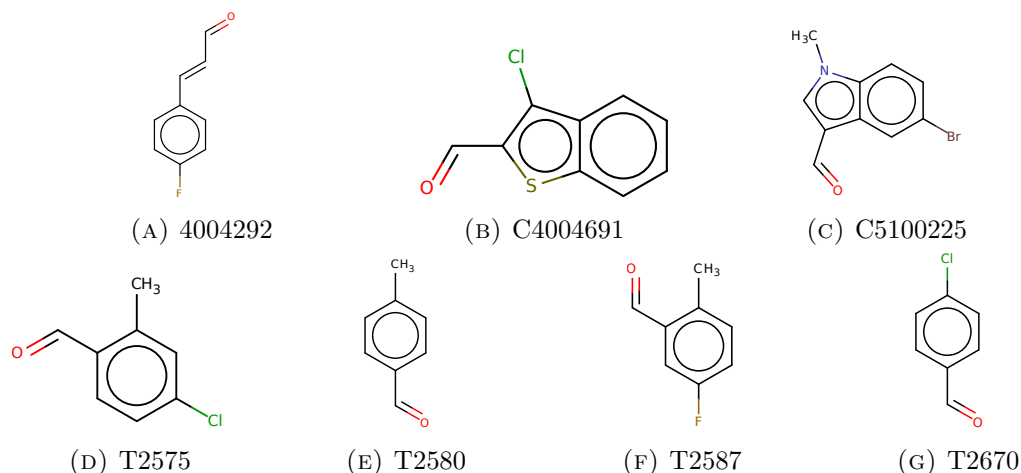


FIGURE 8.1: *Unspecific* or 'safe' building blocks selected for the *reductive amination* series. The carbonyl moiety will disappear during the *reductive amination* reaction.

We can then ask ourselves from what would be a valuable *unspecific* building block library be made? This might be too early to tell, but based on our recent experience we can already point out important trends:

- *unspecific* building blocks are fragment-to-grow and reaction dependent. Indeed, an *unspecific* library designed for *reductive amination* applied to a fragment with an amine moiety won't work for a fragment without such a moiety.
- *unspecific* building blocks should contain at least one aromatic moiety.
- The reactive center and the hydrophobic moiety could be linked with an alkyl chain (i.e. flexible). The distance between the reactive center and the aromatic moiety should vary in order to increase the number of rotors, and thus the conformational space that can be covered by the product. This should increase the chance for the aromatic moiety to accommodate to the receptor constraints.
- The aromatic moiety should be decorated with a wide range of groups at any positions. Some variation should be introduced for the derivatives containing halogens. For instance, chlorine should be prioritized over fluorine due to its aptitude to engage in halogen bonds. Since halogen bonds are specific and highly directional, they could be used to explore dipole partners (e.g. carbonyl) more cheaply than a H-bond donor group, due to lower desolvation penalty.

In the future, *unspecific* building blocks will be flagged in SCUBIDOO in order to provide users with the possibility to screen pre-compiled 'obese' fragments. A new subset of the database will be created, containing every polar building block with extended non-specific moieties.

8.4 Tailored combinatorial synthesis guided by combinatorial growing

Combinatorial synthesis allows one to synthesize as many products as possible, going out from a few initial building blocks. This procedure was common several decades ago [287–289] and massively failed when applied to drug discovery efforts [290]. One of the reason was the challenging task of selecting the best compounds library for synthesis and follow up assays. Such task should rely in computational approaches and, years later, we had the opportunity to try to tackle such challenge. We illustrated this scenario in chapter 7, where we are synthesizing up to 240 molecules with only 58 building blocks (10 A + 48 B). As of right now, no similar scientific work has been reported in the literature. The main challenge in our study was to define where and how to start. In the following paragraphs, I will describe the pillars that I think are crucial to build efficiently large scale applications of tailored combinatorial synthesis.

In a drug discovery project, products originating from combinatorial synthesis, will usually be tested against the target of interest. In a bipartite product context, the main challenge is then to find the combination of building blocks (A and B) which might yield the highest synthesis rate and also the highest experimental hit rate. This task on its own is quite a dilemma. Indeed, one can imagine that providing B suggestions for a single A is straightforward. However, if the same suggested B s are kept and applied to a different A , the resulting products might not preserve the same binding mode, and thus might be more likely to fail in the assays. The selection of the A fragments is thus critical and can be assisted by computational approaches. Three crucial elements ought to be taken into account during the selection process: position of the reactive center, presence of rotors and chemical diversity.

The reactive center (i.e. the chemical feature involved in the organic reaction) must be considered as an anchor point. A fragments should have their anchor points in close vicinity to each other, so the B building blocks are attached in the same region of the binding site, and are thus more likely to preserve the same binding mode (figure 7.7). One way to ensure that could be by filtering the building blocks with a common pharmacophore. For instance, in chapter 7 the A fragments were selected so they have two carbons between the amine and the aromatic moiety (figure 7.8).

Inclusion of rotors in the products can be profitable. According to the bipartite nature of the product, if two bulky moieties are linked together with a flexible chain, this should give more chance to the product to adapt to the receptor constraints. Indeed, the more rotors a molecule has, the more of the conformational space has to be sampled, and thus the more versatile the molecule might. However, such procedure will yield additional

entropic penalties, since every rotor frozen upon binding will count for approximately 1 kcal/mol.

Chemical diversity should be assessed within the pool of A and B fragments to ensure a wide variety amongst the generated products. Doing so might increase the chance of finding hits later on, since the focus is on several residues rather than only a few. This stage could be referred to the “don’t put all your eggs in one basket” stage. Additionally, the pool of B fragments should contain at least 20% of unspecific building blocks (hydrophobic), in order to yield about 20% ‘obese’ products. Doing so could increase the chance to identify hits in the experimental assays.

We are currently witnessing the ‘rise of the automated synthesis machines’ [291]. This technology is not yet within the reach of most of academic institutions. However, its mechanism (bipartite product) could be applied to small scale ligand discovery efforts. For instance, in chapter 4, three derivatives were synthesized for compound Z32501319, thus four building blocks were ordered. In an optimal bipartite product scenario, with four building blocks ($2 A + 2 B$), one could synthesize four products ($2*2$). The difference is even more striking if we start from six building blocks ($3 A + 3 B$), this could yield 9 products ($3*3$), etc... The underlying assumption is that if the A fragments are similar to each other, the reaction conditions should be comparable. Thus the reactions could be done ‘in parallel’ once a robust protocol has been established. Such mechanism might also generate products more cheaply (each building block is used several time for different products synthesis).

Products resulting from tailored combinatorial synthesis offer the advantage to be located in the same region of chemical space, because they are derivatives from the same scaffold (i.e. analogs). This represents a huge opportunity in the QSAR/QSPR field. Such interconnected datasets with associated measured affinity (or reaction yield or solubility) could allow one to create and validate more accurate local models with high discriminative power.

8.5 Exploiting the ‘loser bracket’

We highlighted that virtually screening SCUBIDOO samples allow one to identify promising fragments that can then be further explored. Usually, when visually inspecting the products from a docking run, one quickly discard fragments that contain unfavorable moieties (i.e. unfavorable interactions). However, a fragment often contains one or more chemical features and can thus be decomposed into smaller moieties. Thus, for a fragment, if one moiety happen to be unfavorable even though another one is

compelling, the fragment will be discarded. In some special case, this can be frustrating if we discard an nearly perfect interaction.

For instance in chapter 7, product 18815102 was identified while inspecting the derivative products of building block 4091032. This product was also made using building block 5130031 which contains an acid moiety. This acid feature makes a very compelling interaction between His93 and Trp313, acting like a 'bridge' between those two residues (Figure 8.2). This moiety seems to fit perfectly to this region of the SBP, however the thiophene is clashing with the receptor which is highly unfavorable. Instead of getting rid of this fragment, we decided to give the acid moiety another chance and created the 'loser bracket' which contain all the fragments that were too disappointing to not explore more deeply.

In order to explore the acid moiety lead, all building blocks containing an acid moiety associated an amine or an carbonyl (i.e. compatible with *reductive amination*) were extracted from the Chembridge library. This procedure yielded 139 building blocks and all carbonyl moieties were transformed in their amine "surrogates", as explained in PINGUI (chapter 4. The resulting "surrogates" were then docked into the SBP region of the β_2 AR binding site. Unfortunately, no "surrogates" were found to form the bridge between His93 and Trp313 while interacting with the conserved Asp113. Given the low number of building blocks, this is not surprising. This also stress that one way to improve SCUBIDOO would be to implement more building blocks in the initial library (discussed in SCUBIDOO future). This will increase the chance to find more selective fragments for such interactions, and also increase the number of selected products coming up from the 'loser bracket'. The identification of this acid moiety is now the pillar of a new project aiming at designing selective ligands of the β_2 AR against the β_1 AR (which will be not described in this manuscript).

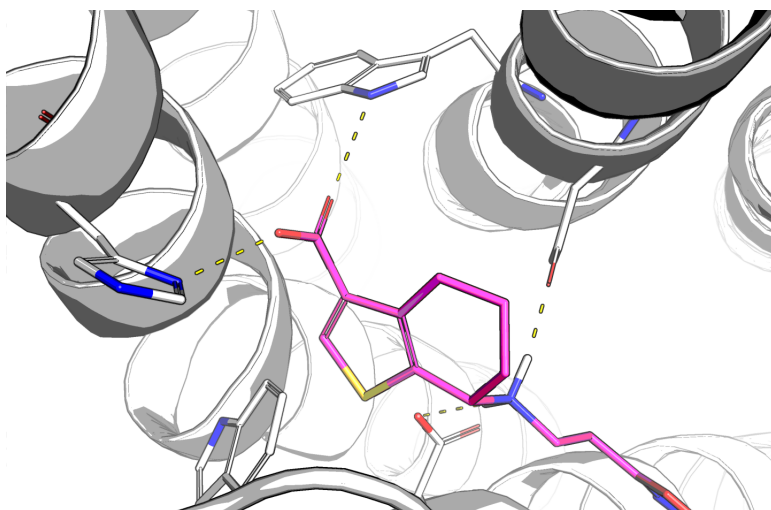


FIGURE 8.2: Fragment falling to the loser bracket

8.6 SCUBIDOO future

While our ongoing endeavors intend to validate the synthetic feasibility of SCUBIDOO products and its applicability to ligand discovery efforts, the next logical step will be to increase its size. Indeed, in chapter 7 we saw some limitations of the current database, when suggesting (too few) *unspecific* building blocks for the *reductive amination* products or our incapacity to rescue compelling chemical moieties coming out from the 'loser bracket'. The former was bypassed by using building blocks from the Taros library and by generating new products with PINGUI. The size of the database can be increased using three parameters: the number of initial building blocks, the number of organic reactions and the number of building blocks allowed for combination.

The initial building blocks library can be expanded by implementing new vendors. At present, SCUBIDOO version 2 implements the Sigma Aldrich building blocks catalog. This library contains about 25'000 building blocks and its combination with the 58 organic reactions yielded more than 300'000'000 virtual products. We were also recently contacted by Enamine (www.enamine.net) who would like to collaborate with us in order to implement their building blocks library (more than 100'000) into SCUBIDOO as well as some in-house organic reactions. Such implementation should yield billions of virtual products. We plan to make this new database version available in early 2017. Ideally, additional building blocks libraries will be implemented regularly in order to provide the users with more ideas to start with.

Hartenfeller et al. [1, 2] did an unprecedented job when they implemented their set of 58 organic reactions. They took great care to select often-used chemical reactions in the pharmaceutical field [248]. But as stated by the authors, this reaction dataset is prone to improvement which can be facilitated by community feedback. We hope to participate in such efforts with our ongoing applications, which already allowed us to learn from the reactions by adding more incompatibility warnings in the database (discussed below). Another way to improve this set will be to increase the number of reactions (which we will do in collaboration with enamine), thus taking into account more complex (multi-step) synthetic routes, less 'trendy' reactions or even implement metabolism reactions.

Although SCUBIDOO was built on a bipartite product philosophy, one could imagine the combination of three building blocks or more in order to increase the sheer size of the database. Indeed, many building blocks contain two reactive features, which thus opens doors to tripartite products (or more). Two major complications might arise from such combinations: combinatorial explosion and over-sized products. Combinatorial explosion should be avoided, because it will generate more chemical noise which will possibly

make the identification of high LE fragments more difficult (i.e. tripartite product syndrome). Secondly, the combination of three building blocks will inevitably yield bigger compounds, beyond the Lipinski frontier. Post filtering is still an option, but the resulting products will still suffer the tripartite product syndrome. SCUBIDOO should focus on helping with the identification of high LE fragments rather than providing users with unexploitable virtual products. Thus, if bigger products are needed, one can turn to PINGUI in order to incrementally build new virtual products. However, tripartite products could allow one to explore 'linking + merging' scenarios with SCUBIDOO (chapter 2). Indeed, such products could be the fruits resulting from a linking strategy, where the same building block is grown twice (each time in a different directions) and the resulting products are then merged as illustrated in figure 2.6.

One of the next challenges for SCUBIDOO will be the implementation of an efficient similarity search tool. As mentioned above, with the expected growth of the database, one will need to be able to parse billions of compounds within a few seconds. With the current ligand-based techniques, this is unrealistic. In order to circumvent the problem, one would have to rely on efficient organization of the data. The key will be to perform the search at a building block level, in order to narrow down the comparison to only a few thousand compounds. Such searches are in the range of seconds on a single CPU. We already took small steps in this direction. For instance PINGUI allows one to perform similarity search against the ChemBridge building blocks library (i.e. the one used for SCUBIDOO creation). One can then identify similar building blocks for a given target, and download their derivatives (i.e. even more analogs). Furthermore, PINGUI provides a tool to deconstruct molecules based on chemical reactions. Such deconstruction predicts the building blocks that were likely utilized for the synthesis of a given target. With those two combined tools, users should be able to perform similarity searches and identify resulting analogs within a few mouse clicks.

The combinatorial synthesis project yielded 127 products with soon associated activity against the β_2 AR, but also synthesis yield and purity. We plan to make these data and the one associated with the future project publicly available. Ideally, we would like to invite the scientific community to a QSAR and a QSPR challenge, similar to the solubility challenge [292, 293]. The goal will be to predict the affinities against the β_2 AR and the synthetic yields for 20% of the products (external test set), by learning from the remaining 80% (training set). The external test set will remain unknown until the end of the challenge. Other properties might be measured afterwards, such as solubility or logP and should be implemented in the challenge too. Measured affinities will also be utilized in order to create subsets specific to certain targets. Indeed, if a building block is found to be active against a given target, all its derivatives as well as close analogs should be 'packed' together in order to facilitate the probing of the target's 'active' chemical space.

Such building blocks could be flagged in the database in order to provide users with a starting kit for relevant pharmaceutical targets.

Further improvements are needed in the direction of simplifying the synthesis process. Additional warnings are needed to guide the users to select more reliable products. For instance, in chapter 7 for the *amide formation* series, the initial *A* fragments contained an amine and an acid moiety (figure 7.7). Under *Schotten-Baumann amide* reaction conditions, such building blocks are highly likely to react with themselves. Thus, one first needs to protect the amine moiety with a Boc group, apply the *amide formation* reaction and finally deprotect in order to obtain the product. In order to make such processes more obvious, additional warnings combined with a protecting group library will be added to the database.

As mentioned earlier, 'obese' extensions will be flagged in order to provide the user with the possibility to screen products with a *specific* head and an *unspecific* tail. Thus, for each polar building block, a new derivative library will be created containing only *unspecific* extensions. An additional subset which contains a representative portion of all those derivatives will also be made available for download. Ultimately, one could also think about synthesizing such a subset and make it available for experimental screening, in order to increase the chances to identify fragment with favorable LE.

With the expected increased size of the database, more filters and user options, the hardware should keep up the pace so SCUBIDOO can always offer its service at reasonable speed. I was recently awarded with an Amazon grant for SCUBIDOO (AWS cloud credits for research). This grant will allow us to transition to faster and more scalable hardware. It should facilitate the deployment of SCUBIDOO 2, which will contain more than 300 M virtual products but also the next versions that will likely contain billions of chemical entities.

Last, but not the least, what does SCUBIDOO aspire to be? The core concept, which is its accessibility, should remain the same: any user with internet-access should be able to benefit from the database. Further tools will be implemented in order to allow one to process more requests from the server side. In addition, with the first large application we ambition to demonstrate how SCUBIDOO could be applied in synergy with automated synthesis robots, thus illustrating how we could quickly come up with hundreds of suggestions and synthesize them when starting a new drug-discovery effort. Follow up assays will hopefully validate our computational assumptions. Upon successful completion, this collaboration will then apply the same strategy to other challenging targets in the field of neglected diseases (Chagas' disease, malaria, tuberculosis), Parkinson's disease or Alzheimer's disease. We would also like SCUBIDOO to support open source

initiatives in the field of drug discovery, such as *open source pharma* (OSP). Such initiatives aim at delivering drugs at affordable prices to anyone. In this context, SCUBIDOO could help to explore *cheap chemical space* where one could extract diverse and affordable products. This will allow one to start a new drug discovery effort focusing on a neglected target, with minimal investment.

Bibliography

- [1] Markus Hartenfeller, Martin Eberle, Peter Meier, Cristina Nieto-Oberhuber, Karl-Heinz Altmann, Gisbert Schneider, Edgar Jacoby, and Steffen Renner. A collection of robust organic synthesis reactions for in silico molecule design. *J. Chem. Inf. Model.*, 51(12):3093–3098, 2011. doi: 10.1021/ci200379p.
- [2] Markus Hartenfeller, Martin Eberle, Peter Meier, Cristina Nieto-Oberhuber, Karl Heinz Altmann, Gisbert Schneider, Edgar Jacoby, and Steffen Renner. Probing the bioactivity-relevant chemical space of robust reactions and common molecular building blocks. *J. Chem. Inf. Model.*, 52:1167–1178, 2012.
- [3] Michael M. Hann. Molecular obesity, potency and other addictions in drug discovery. *Med. Chem. Commun.*, 2:349–355, 2011.
- [4] Craigh W. Lindsley. New statistics on the cost of new drug development and the trouble with cns drugs. *ACS chemical neuroscience*, (5):1142–1142, 2014.
- [5] Roger Gunn and Ilan Rabiner. Making drug development visible - and viable. *Drug Discov. Today*, 19(1):1–3, 2014.
- [6] Philip J Hajduk and Jonathan Greer. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug. Discov.*, 6(3):211–219, 2007.
- [7] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and developmental settings. *Advanced Drug Deliv. Rev.*, 23:3–25, 1997.
- [8] Christopher a. Lipinski. Lead- and drug-like compounds: The rule-of-five revolution. *Drug Discovery Today: Technologies*, 1(4):337–341, 2004.
- [9] Haijun Chen, Xiaobin Zhou, Ailan Wang, Yunquan Zheng, Yu Gao, and Jia Zhou. Evolutions in fragment-based drug design: the deconstruction–reconstruction approach. *Drug Discov. Today*, 20(1):105–113, 2015.

- [10] Christopher W. Murray, Marcel L. Verdonk, and David C. Rees. Experiences in fragment-based drug discovery. *Trends Pharmacol. Sci.*, 33(5):224–232, 2012.
- [11] Shuker Suzanne B. Hjdk Philip J. Meadows Robert P. Fesik, Stephen W. Discovering high-affinity ligands for proteins: Sar by nmr. *Science*, 274(November), 1996.
- [12] Miles Congreve, Robin Carr, Chris Murray, and Harren Jhoti. A ‘rule of three’ for fragment-based lead discovery? *Drug Discov. Today*, 8(19):876–877, 2003.
- [13] I. D. Kuntz, K. Chen, K. A. Sharp, and P. A. Kollman. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U. S. A.*, 96:9997–10002, 1999.
- [14] Andrew L. Hopkins, Colin R. Groom, and Alexander Alex. Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today*, 9:430–431, 2004.
- [15] Christopher W Murray, Daniel a Erlanson, Andrew L Hopkins, Gyo M Keseru, Paul D Leeson, David C Rees, Charles H Reynolds, and Nicola J Richmond. Validity of ligand efficiency metrics. *ACS Med. Chem. Lett.*, 5:616–618, 2014.
- [16] Michael M. Hann, Andrew R. Leach, and G. Harper. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.*, 41:856–864, 2001.
- [17] Dima Kozakov, David R. Hall, Sefan Jehle, Lingqi Luo, Stefan O. Ochiana, Elizabeth V. Jones, Michael Pollastri, Karen N. Allen, Adrian Whitty, and Sandor Vajda. Ligand deconstruction: Why some fragment binding positions are conserved and others are not. *Proc. Natl. Acad. Sci. U. S. A.*, page 201501567, 2015.
- [18] Ansgar Schuffenhauer, Simon Ruedisser, Andreas L Marzinzik, Wolfgang Jahnke, Marcel Blommers, Paul Selzer, and Edgar Jacoby. Library design for fragment based screening. *Curr Top Med Chem*, 5(8):751–762, 2005.
- [19] Kathryn Loving, Ian Alberts, and Woody Sherman. Computational approaches for fragment-based and de novo design. *Curr Top Med Chem*, 10(1):14–32, 2010.
- [20] Gideon Bollag, James Tsai, Jiazhong Zhang, Chao Zhang, Prabha Ibrahim, Keith Nolop, and Peter Hirth. Vemurafenib: the first drug approved for braf-mutant cancer. *Nat Rev Drug Discov*, 11:873–886, 2012.
- [21] Kristin E. D. Coan and Brian K. Shoichet. Stoichiometry and physical chemistry of promiscuous aggregate-based inhibitors. *J. Am. Chem. Soc.*, 130(29):9606–9612, 2008. doi: 10.1021/ja802977h.

- [22] Diane Joseph-McCarthy, Arthur J. Campbell, Gunther Kern, and Demetri Moustakas. Fragment-based lead discovery and design. *J. Chem. Inf. Model.*, 54(3): 693–704, 2014.
- [23] Ben J. Davis and Daniel a. Erlanson. Learning from our mistakes: The ‘unknown knowns’ in fragment screening. *Bioorg. Med. Chem. Lett.*, 23(10):2844–2852, 2013.
- [24] Gyorgy M Keseru, Daniel Andrew Erlanson, Gyorgy G Ferenczy, Michael M Hann, Christopher W Murray, and Stephen D Pickett. Design principles for fragment libraries—maximizing the value of learnings from pharma fragment based drug discovery (fbdd) programs for use in academia. *J. Med. Chem.*, 2016.
- [25] Johannes Schiebel, Nedyalka Radeva, Helene Köster, Alexander Metz, Timo Krotzky, Maren Kuhnert, Wibke E Diederich, Andreas Heine, Lars Neumann, Cedric Atmanene, et al. One question, multiple answers: Biochemical and biophysical screening methods retrieve deviating fragment hit lists. *ChemMedChem*, 10(9):1511–1521, 2015.
- [26] Johannes Schiebel, Nedyalka Radeva, Stefan G Krimmer, Xiaojie Wang, Martin Stieler, Frederik R Ehrmann, Kan Fu, Alexander Metz, Fransiska U Huschmann, Manfred S Weiss, et al. Six biophysical screening methods miss a large proportion of crystallographically discovered fragment hits: A case study. *ACS Chem. Biol.*, 2016.
- [27] Helene Köster, Tobias Craan, Sascha Brass, Christian Herhaus, Matthias Zentgraf, Lars Neumann, Andreas Heine, and Gerhard Klebe. A small nonrule of 3 compatible fragment library provides high hit rate of endothiapepsin crystal structures with various fragment chemotypes. *J. Med. Chem.*, 54, 2011.
- [28] Mark Ashton, John Barnard, Florence Casset, Michael Charlton, Geoffrey Downs, Dominique Gorse, John Holliday, Roger Lahana, and Peter Willett. Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quantitative Structure-Activity Relationships*, 21(6):598–604, 2002.
- [29] M. J. McGregor and P. V. Pallai. Clustering of large database of compounds: Using mdl keys as structural descriptors. *J. Chem. Inf. Comput. Sci.*, 37(3):443–448, 1997. doi: 10.1021/ci960151e.
- [30] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.
- [31] Guillaume Chauvet. Stratified balanced sampling. *Surv. Methodol.*, 35(1):115–119, 2009.

- [32] Anton Grafstrom. *BalancedSampling: Balanced and spatially balanced sampling* version 1.4. URL <http://www.antongrafstrom.se/balancedsampling>. (accessed Nov 15, 2014).
- [33] Jean-claude Deville, Blaise Pascal, and Yves Tille Groupe. Efficient balanced sampling: The cube method. *Biometrika*, 91(4):893–912, 2004.
- [34] Caren Hasler and Yves Tillé. Fast balanced sampling for highly stratified population. *Comput. Stat. Data Anal.*, 74:81–94, 2014. doi: 10.1016/j.csda.2013.12.005.
- [35] Chunquan Sheng and Zhang Wannian. Fragment informatics and computational fragment-based drug design: An overview and update. *Med Res Rev*, (6):1292–1327, 2012.
- [36] Gregg Siegal, Eiso Ab, and Jan Schultz. Integration of fragment screening and library design. *Drug Discov. Today*, 12(23-24):1032–1039, 2007.
- [37] Gergely M. Makara. On sampling of fragment space. *J. Med. Chem.*, 50(14):3214–3221, 2007.
- [38] Li Di and Edward H. Kerns. Biological assay challenges from compound solubility: strategies for bioassay optimization. *Drug Discov. Today*, 11(9-10):446–451, 2006.
- [39] Christopher a. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*, volume 64, pages 4–17. Elsevier B.V., 2012.
- [40] Junmei Wang and Tingjun Hou. Recent advances on aqueous solubility prediction. *Combinatorial chemistry & high throughput screening*, 14(5):328–338, 2011.
- [41] Florent Chevillard, David Lagorce, Christelle Reynès, Bruno O Villoutreix, Philippe Vayer, and Maria Miteva. Multimodel Protocol Based on Chemical Similarity In silico Prediction of Aqueous Solubility : A Multimodel Protocol Based on Chemical Similarity. *Mol. Pharm.*, 9(11):3127–3135, 2012.
- [42] John B Jordan, Leszek Poppe, Xiaoyang Xia, Alan C Cheng, Yax Sun, Klaus Michelsen, Heather Eastwood, Paul D Schnier, Thomas Nixey, and Wenge Zhong. Fragment based drug discovery: practical implementation based on 19f nmr spectroscopy. *J. Med. Chem.*, 55(2):678–687, 2012.
- [43] Clifford T Gee, Edward J Koleski, and William CK Pomerantz. Fragment screening and druggability assessment for the cbp/p300 kix domain through protein-observed 19f nmr spectroscopy. *Angew. Chem. Int. Ed. Engl.*, 54(12):3735–3739, 2015.

- [44] Miguel Garavís, Blanca López-Méndez, Alvaro Somoza, Julen Oyarzabal, Claudio Dalvit, Alfredo Villasante, Ramón Campos-Olivas, and Carlos González. Discovery of selective ligands for telomeric rna g-quadruplexes (terra) through 19f-nmr based fragment screening. *ACS chem. biol.*, 9(7):1559–1566, 2014.
- [45] Marina Veronesi, Elisa Romeo, Chiara Lambruschini, Daniele Piomelli, Tiziano Bandiera, Rita Scarpelli, Gianpiero Garau, and Claudio Dalvit. Fluorine nmr-based screening on cell membrane extracts. *ChemMedChem*, 9(2):286–289, 2014.
- [46] Jonathan B. Baell and Georgina A. Holloway. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.*, 53(7):2719–2740, 2010. doi: 10.1021/jm901137j.
- [47] Konstantin Arnold, Lorenza Bordoli, Jürgen Kopp, and Torsten Schwede. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201, 2006. doi: 10.1093/bioinformatics/bti770.
- [48] Khaled M Elokely and Robert J Doerksen. Docking challenge: protein sampling and molecular docking performance. *J. Chem. Inf. Model.*, 53(8):1934–1945, 2013.
- [49] Peter Kolb and Amedeo Caffisch. Automatic and efficient decomposition of two-dimensional structures of small molecules for fragment-based high-throughput docking. *J. Med. Chem.*, 49:7384–7392, 2006.
- [50] Alasdair TR Laurie and Richard M Jackson. Q-sitefinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics*, 21(9):1908–1916, 2005.
- [51] Manfred Hendlich, Friedrich Rippmann, and Gerhard Barnickel. Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling*, 15(6):359–363, 1997.
- [52] Peter Schmidtke, Catherine Souaille, Frédéric Estienne, Nicolas Baurin, and Romano T Kroemer. Large-scale comparison of four binding site detection algorithms. *J. Chem. Inf. Model.*, 50(12):2191–2200, 2010.
- [53] Peter Schmidtke, Vincent Le Guilloux, Julien Maupetit, and Pierre Tufféry. Fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic acids research*, 38(suppl 2):W582–W589, 2010.
- [54] Andrea Volkamer, Axel Griewel, Thomas Grombacher, and Matthias Rarey. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.*, 50(11):2041–2052, 2010.

- [55] Andrea Volkamer, Daniel Kuhn, Friedrich Rippmann, and Matthias Rarey. Dogsitescorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics*, 28(15):2074–2075, 2012.
- [56] Andrea Volkamer, Daniel Kuhn, Thomas Grombacher, Friedrich Rippmann, and Matthias Rarey. Combining global and local measures for structure-based drug-gability predictions. *J. Chem. Inf. Model.*, 52(2):360–372, 2012.
- [57] Paul C. D. Hawkins, A. Geoffrey Skillman, Gregory L. Warren, Benjamin A. Ellingson, and Matthew T. Stahl. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and cambridge structural database. *J. Chem. Inf. Model.*, 50(4):572–584, 2010.
- [58] Mark McGann. FRED pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.*, (1):578–596, 2011.
- [59] J. Sadowski, J. Gasteiger, and Klebe G. Comparison of automatic three-dimensional model builders using 639 x-ray structures. *J. Chem. Inf. Comput. Sci.*, 34:1000–1008, 1994.
- [60] Peter Kolb, Marco Cecchini, Danzhi Huang, and Amedeo Caflisch. Fragment-based high-throughput docking. In Juan Alvarez and Brian K. Shoichet, editors, *Virtual Screening in Drug Discovery*, pages 349–378. CRC Press, Boca Rato, FL, USA, 2005.
- [61] Holger Claußen, Christian Buning, Matthias Rarey, and Thomas Lengauer. FlexE: Efficient molecular docking considering protein structure variations. *Algorithmica*, 308:377–395, 2001.
- [62] Todd J. A. Ewing, Shakino Makino, A. Geoffrey Skillman, and Irwin D. Kuntz. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Design*, 15(5):411–428, 2001.
- [63] Matthias Rarey, Bernd Kramer, Thomas Lengauer, and Gerhard Klebe. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261:470–489, 1996.
- [64] Garrett M. Morris, David S. Goodsell, Robert S. Halliday, Ruth Huey, William E. Hart, Richard K. Belew, and Arthur J. Olson. Automated docking using a Lamarckian Genetic Algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19(14):1639–1662, 1998.
- [65] Nicolas Budin, Shaheen Ahmed, Nicolas Majeux, and Amedeo Caflisch. An evolutionary approach for structure-based design of natural and non-natural peptidic ligands. *Comb. Chem. High Throughput Screen.*, 4:695–707, 2001.

- [66] Nicolas Budin, Nicolas Majeux, Catherine Tenette-Souaille, and Amedeo Caflisch. Structure-based ligand design by a build-up approach and genetic algorithm search in conformational space. *J. Comput. Chem.*, 22(16):1956–1970, 2001.
- [67] Nicolas Budin, Nicolas Majeux, and Amedeo Caflisch. Fragment-based flexible ligand docking by evolutionary optimization. *Biol. Chem.*, 382:1365–1372, 2001.
- [68] Michael Thormann and Miquel Pons. Massive docking of flexible ligands using environmental niches in parallelized genetic algorithms. *J. Comput. Chem.*, 22(16):1971–1982, 2001.
- [69] Fabian Dey and Amedeo Caflisch. Fragment-based de novo ligand design by multi-objective evolutionary optimization. *J. Chem. Inf. Model.*, 48:679–690, 2008.
- [70] Gareth Jones, Peter Willett, and Robert C. Glen. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.*, 245:43–53, 1995.
- [71] C. M. Oshiro, I. D. Kuntz, and J. S. Dixon. Flexible ligand docking using a genetic algorithm. *J. Comput.-Aided Mol. Design*, 9:113–130, 1995.
- [72] C. M. Venkatachalam, X. Jiang, T. Oldfield, and M. Waldman. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Model.*, 21:289–307, 2003.
- [73] R.S. DeWitte and E.I. Shakhnovich. SMOG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.*, 118:11733–11744, 1996.
- [74] R.S. DeWitte, A.V. Ishchenko, and E.I. Shakhnovich. SMOG: de novo design method based on simple, fast, and accurate free energy estimates. 2. Case studies in molecular design. *J. Am. Chem. Soc.*, 119:4608–4617, 1997.
- [75] D B Kitchen, H Decornez, J R Furr, and J Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug. Discov.*, 3:935–949, 2004.
- [76] R. D. Taylor, P. J. Jewsbury, and J. W. Essex. A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Design*, 16(3):151–166, 2002.
- [77] Nicolas Majeux, Marco Scarsi, Joannis Apostolakis, Claus Ehrhardt, and Amedeo Caflisch. Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins: Struct., Funct., Bioinf.*, 37:88–105, 1999.

- [78] Nicolas Majeux, Marco Scarsi, and Amedeo Caflisch. Efficient electrostatic solvation model for protein-docking. *Proteins: Struct., Funct., Bioinf.*, 42:256–268, 2001.
- [79] P.-G. Mailliot. *Graphics Gems*, page 498. Academic Press, London, 1996.
- [80] Mark R. McGann, Harold R. Almond, Anthony Nicholls, J. Andrew Grant, and Frank K. Brown. Gaussian docking functions. *Biopolymers*, 68(1):76–90, 2003.
- [81] Mark McGann. Fred and hybrid docking performance on standardized datasets. *J. Comput. Aided Mol. Des.*, 26(8):897–906, 2012.
- [82] Georgia B. McGaughey, Robert P. Sheridan, Christopher I. Bayly, J. Chris Culberston, Constantine Kreatsoulas, Stacey Lindsley, Vladimir Maiorov, Jean Francois Truchon, and Wendy D. Cornell. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.*, 47(4):1504–1519, 2007.
- [83] Irwin D. Kuntz, E. C. Meng, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161:269–288, 1982.
- [84] Ryan G. Coleman, Michael Carchia, Teague Sterling, John J. Irwin, and Brian K. Shoichet. Ligand pose and orientational sampling in molecular docking. *PLoS ONE*, 8(10):e75992, 10 2013.
- [85] Volker Schnecke and Leslie Kuhn. Virtual screening with solvation and ligand-induced complementarity. *Persp. Drug Disc. Des.*, 20:171–190, 2000.
- [86] B Kramer, M Rarey, and T Lengauer. Evaluation of the flexx incremental construction algorithm for protein-ligand docking. *Proteins*, 37(2):228–41, 1999.
- [87] NM OpenEye Scientific Software, Santa Fe. *SZYBKI* 1.8.0.2, . URL <http://www.eyesopen.com>. (accessed Nov 15, 2014).
- [88] Virginie Y Martiny, Pablo Carbonell, Florent Chevillard, Gautier Moroy, B Arnaud, Philippe Vayer, Bruno O Villoutreix, and Maria A Miteva. Integrated structure- and ligand-based in silico approach to predict inhibition of cytochrome p450 2d6. *Bioinformatics*, pages 1–8, 2015.
- [89] Hans Joachim Böhm. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput.-Aided Mol. Design*, 6(6):593–606, 1992.
- [90] Matthew D. Eldridge, Christopher W. Murray, Timothy R. Auton, Gaia V. Paolini, and Roger P. Mee. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Design*, 11(5):425–445, 1997.

- [91] Didier Rognan, Sanne Lise Lauemøller, Arne Holm, Søren Buus, and Vincenzo Tschinke. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class i major histocompatibility proteins. *J. Med. Chem.*, 42(22):4650–4658, 1999.
- [92] Doree Sitkoff, Kim A Sharp, and Barry Honig. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, 98(7):1978–1988, 1994.
- [93] Shuanghong Huo, Junmei Wang, Piotr Cieplak, Peter A Kollman, and Irwin D Kuntz. Molecular dynamics and free energy analyses of cathepsin d-inhibitor interactions: insight into structure-based ligand design. *J. Med. Chem.*, 45(7):1412–1419, 2002.
- [94] Ingo Muegge. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. *Persp. Drug Discov. Des.*, 20:99–114, 2000.
- [95] Ingo Muegge. Effect of ligand volume correction on pmf scoring. *J. Comput. Chem.*, 22(4):418–425, 2001.
- [96] Ingo Muegge and Yvonne C Martin. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.*, 42(5):791–804, 1999.
- [97] Hans FG Velec, Holger Gohlke, and Gerhard Klebe. Drugscorecsd knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.*, 48(20):6296–6303, 2005.
- [98] Holger Gohlke, Manfred Hendlich, and Gerhard Klebe. Predicting binding modes, binding affinities and hot spots’ for protein-ligand complexes using a knowledge-based scoring function. *Perspectives in Drug Discovery and Design*, 20(1):115–144, 2000.
- [99] Holger Gohlke, Manfred Hendlich, and Gerhard Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, 295(2):337–356, 2000.
- [100] Alan P Graves, Devleena M Shivakumar, Sarah E Boyce, Matthew P Jacobson, David A Case, and Brian K Shoichet. Rescoring docking hit lists for model cavity sites: predictions and experimental testing. *J. Mol. Biol.*, 377(3):914–934, 2008.
- [101] Gerd Neudert and Gerhard Klebe. Dsx: a knowledge-based scoring function for the assessment of protein–ligand complexes. *J. Chem. Inf. Model.*, 51(10):2731–2745, 2011.

- [102] Nadine Schneider, Gudrun Lange, Sally Hindle, Robert Klein, and Matthias Rarey. A consistent description of hydrogen bond and dehydration energies in protein–ligand complexes: methods behind the hyde scoring function. *J. Chem. Inf. Model.*, 27(1):15–29, 2013.
- [103] A. Miranker and M. Karplus. Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins: Struct., Funct., Bioinf.*, 11:29–34, 1991.
- [104] M L Verdonk, J C Cole, and R Taylor. Superstar: a knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.*, 289(4):1093–1108, 1999.
- [105] Marcel L. Verdonk, Jason C. Cole, Paul Watson, Valerie J. Gillet, and Peter Willett. SuperStar: Improved knowledge-based interaction fields for protein binding sites. *J. Mol. Biol.*, 307(3):841–859, 2001.
- [106] Marcel L Verdonk, Ilenia Giangreco, Richard J Hall, Oliver Korb, Paul N Mortenson, and Christopher W Murray. Docking performance of fragments and druglike compounds. *J. Med. Chem.*, 54(15):5422–5431, 2011.
- [107] Duncan E Scott, Anthony G Coyne, Sean A Hudson, and Chris Abell. Fragment-based approaches in drug discovery and chemical biology. *Biochemistry*, 51(25):4990–5003, 2012.
- [108] Andrew Anighoro, Dagmar Stumpfe, Kathrin Heikamp, Kristin Beebe, Leonard M Neckers, Juergen Bajorath, and Giulio Rastelli. Computational polypharmacology analysis of the heat shock protein 90 interactome. *J. Chem. Inf. Model.*, 55(3):676–686, 2015.
- [109] Giulio Rastelli and Luca Pinzi. Computational polypharmacology comes of age. *Front Pharmacol*, 6, 2015.
- [110] Philip Rawlins. Current trends in label-free technologies. *Drug Discovery World*, 11(4):17–26, 2010.
- [111] S. Bartoli, C. Fincham, and D. Fattori. Fragment-based drug design: combining philosophy with technology. *Curr. Opin. Drug Discov. Devel.*, 10:422–429, 2010.
- [112] Alessio Ciulli, Glyn Williams, Alison G. Smith, Tom L. Blundell, and Chris Abell. Probing hot spots at protein-ligand binding sites: A fragment-based approach using biophysical methods. *J. Med. Chem.*, 49(16):4992–5000, 2006.
- [113] Philip J Hajduk, Jeffrey R Huth, and Stephen W Fesik. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.*, 48:2518–2525, 2005.

- [114] T Neumann, HD Junker, K Schmidt, and R Sekul. Spr-based fragment screening: advantages and applications. *Curr. Top. Med. Chem.*, 7(16):1630–1642, 2007.
- [115] U Helena Danielson. Fragment library screening and lead characterization using spr biosensors. *Curr. Top. Med. Chem.*, 9(18):1725–1735, 2009.
- [116] Samantha Perspicace, David Banner, Jörg Benz, Francis Müller, Daniel Schlatter, and Walter Huber. Fragment-based screening using surface plasmon resonance technology. *Journal of biomolecular screening*, 14(4):337–349, 2009.
- [117] Laurent Hoffer, Jean-Paul Renaud, and Dragos Horvath. Fragment-based drug design: computational and experimental state of the art. *Combinatorial chemistry & high throughput screening*, 14(6):500–520, 2011.
- [118] a. Kumar, a. Voet, and K.Y.J. Zhang. Fragment based drug design: From experimental to computational approaches. *Current Medicinal Chemistry*, 19(30):5128–5147, 2012.
- [119] Iva Navratilova and Andrew L Hopkins. Fragment screening by surface plasmon resonance. *ACS Med. Chem. Lett.*, 1(1):44–48, 2010.
- [120] Tonia Aristotelous, Seungkirl Ahn, Arun K Shukla, Sylwia Gawron, Maria F Sassano, Alem W Kahsai, Laura M Wingler, Xiao Zhu, Prachi Tripathi-Shukla, Xi-Ping Huang, et al. Discovery of β_2 adrenergic receptor ligands using biosensor fragment screening of tagged wild-type receptor. *ACS Med. Chem. Lett.*, 4(10):1005–1010, 2013.
- [121] Andreas Kuglstatter, Martin Stahl, Jens-Uwe Peters, Walter Huber, Martine Stihle, Daniel Schlatter, Jörg Benz, Armin Ruf, Doris Roth, Thilo Enderle, et al. Tyramine fragment binding to bace-1. *Bioorg. Med. Chem. Lett.*, 18(4):1304–1307, 2008.
- [122] Helena Nordstrom, Thomas Gossas, Markku Hamalainen, Per Kallblad, Susanne Nystrom, Hans Wallberg, and U Helena Danielson. Identification of mmp-12 inhibitors by using biosensor-based screening of a fragment library. *J. Med. Chem.*, 51(12):3449–3459, 2008.
- [123] Markku D Hämäläinen, Andrei Zhukov, Maria Ivarsson, Tomas Fex, Johan Gottfries, Robert Karlsson, and Magnus Björnsne. Label-free primary screening and affinity ranking of fragment libraries using parallel analysis of protein panels. *Journal of biomolecular screening*, 13(3):202–209, 2008.
- [124] John E Ladbury. Counting the calories to stay in the groove. *Structure*, 3(7):635–639, 1995.

- [125] John E Ladbury and Babur Z Chowdhry. Sensing the heat: the application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions. *Chemistry & biology*, 3(10):791–801, 1996.
- [126] John E Ladbury. Isothermal titration calorimetry: application to structure-based drug design. *Thermochimica acta*, 380(2):209–215, 2001.
- [127] Andrew D Scott, Chris Phillips, Alexander Alex, Maria Flocco, Andrew Bent, Amy Randall, Ronan O’Brien, Luminita Damian, and Lyn H Jones. Thermodynamic optimisation in drug discovery: a case study using carbonic anhydrase inhibitors. *ChemMedChem*, 4(12):1985–1989, 2009.
- [128] John E Ladbury, Gerhard Klebe, and Ernesto Freire. Adding calorimetric data to decision making in lead discovery: a hot tip. *Nature Reviews Drug Discovery*, 9(1):23–27, 2010.
- [129] Ewald Edink, Prakash Rucktooa, Kim Retra, Atilla Akdemir, Tariq Nahar, Obbe Zuiderveld, René van Elk, Elwin Janssen, Pim van Nierop, Jacqueline van Muijlwijk-Koezen, et al. Fragment growing induces conformational changes in acetylcholine-binding protein: a structural and thermodynamic analysis. *J. Am. Chem. Soc.*, 133(14):5363–5371, 2011.
- [130] Matthew A Cooper. Label-free screening of bio-molecular interactions. *Analytical and bioanalytical chemistry*, 377(5):834–842, 2003.
- [131] James K Kranz and Celine Schalk-Hihi. Protein thermal shifts to identify low molecular weight fragments. *Methods Enzymol*, 493:277–298, 2011.
- [132] Maxwell D Cummings, Michael A Farnum, and Marina I Nelen. Universal screening methods and applications of thermofluor®. *Journal of biomolecular screening*, 11(7):854–863, 2006.
- [133] Mei-Chu Lo, Ann Aulabaugh, Guixian Jin, Rebecca Cowling, Jonathan Bard, Michael Malamas, and George Ellestad. Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Analytical biochemistry*, 332(1):153–159, 2004.
- [134] Michael W Pantoliano, Eugene C Petrella, Joseph D Kwasnoski, Victor S Lobanov, James Myslik, Edward Graf, Ted Carver, Eric Asel, Barry A Springer, Pamela Lane, et al. High-density miniaturized thermal shift assays as a general strategy for drug discovery. *Journal of biomolecular screening*, 6(6):429–440, 2001.
- [135] Masoud Vedadi, Frank H Niesen, Abdellah Allali-Hassani, Oleg Y Fedorov, Patrick J Finerty, Gregory A Wasney, Ron Yeung, Cheryl Arrowsmith, Linda J

- Ball, Helena Berglund, et al. Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. *Proceedings of the National Academy of Sciences*, 103(43):15835–15840, 2006.
- [136] Nicolas Basse, Joel L Kaar, Giovanni Settanni, Andreas C Joerger, Trevor J Rutherford, and Alan R Fersht. Toward the rational design of p53-stabilizing drugs: probing the surface of the oncogenic y220c mutant. *Chemistry & biology*, 17(1):46–56, 2010.
- [137] Susan L McGovern, Emilia Caselli, Nikolaus Grigorieff, and Brian K Shoichet. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.*, 45(8):1712–1722, 2002.
- [138] Gilbert M Rishton. Nonleadlikeness and leadlikeness in biochemical screening. *Drug discovery today*, 8(2):86–96, 2003.
- [139] Gilbert M Rishton. Reactive compounds and in vitro false positives in hts. *Drug Discov. Today*, 2(9):382–384, 1997.
- [140] Robert Godemann, James Madden, Joachim Kramer, Myron Smith, Ulrike Fritz, Thomas Hesterkamp, John Barker, Sabine Hoppner, David Hallett, Andrea Cesura, et al. Fragment-based discovery of bace1 inhibitors using functional assays. *Biochemistry*, 48(45):10743–10751, 2009.
- [141] John J Barker, Oliver Barker, Roberto Boggio, Viddhata Chauhan, Robert KY Cheng, Vincent Corden, Stephen M Courtney, Neil Edwards, Virginie M Falque, Fulvia Fusar, et al. Fragment-based identification of hsp90 inhibitors. *ChemMed-Chem*, 4(6):963–966, 2009.
- [142] M Cris Silva-Santisteban, Isaac M Westwood, Kathy Boxall, Nathan Brown, Sam Peacock, Craig McAndrew, Elaine Barrie, Meirion Richards, Amin Mirza, Antony W Oliver, et al. Fragment-based screening maps inhibitor interactions in the atp-binding site of checkpoint kinase 2. *PLoS ONE*, 8(6):e65689, 2013.
- [143] James Tsai, John T Lee, Weiru Wang, Jiazhong Zhang, Hanna Cho, Shumeye Mamo, Ryan Bremer, Sam Gillette, Jun Kong, Nikolas K Haass, Katrin Sproesser, Ling Li, Keiran S M Smalley, Daniel Fong, Yong-Liang Zhu, Adhirai Marimuthu, Hoa Nguyen, Billy Lam, Jennifer Liu, Ivana Cheung, Julie Rice, Yoshihisa Suzuki, Catherine Luu, Calvin Settachatgul, Rafe Shellooe, John Cantwell, Sung-Hou Kim, Joseph Schlessinger, Kam Y J Zhang, Brian L West, Ben Powell, Gaston Habets, Chao Zhang, Prabha N Ibrahim, Peter Hirth, Dean R Artis, Meenhard Herlyn, and Gideon Bollag. Discovery of a selective inhibitor of oncogenic b-raf kinase with

- potent antimelanoma activity. *Proc. Natl. Acad. Sci. U. S. A.*, 105(8):3041–3046, 2008.
- [144] Valerie Vivat Hannah, C Atmanene, D Zeyer, A Van Dorselaer, and Sarah Sanglier-Cianférani. Native ms: an’esi’way to support structure-and fragment-based drug discovery. *Future medicinal chemistry*, 2(1):35–50, 2010.
 - [145] Steven A Hofstadler and Kristin A Sannes-Lowery. Applications of esi-ms in drug discovery: interrogation of noncovalent complexes. *Nat. Rev. Drug. Discov.*, 5(7): 585–595, 2006.
 - [146] Nyssa Drinkwater, Hoan Vu, K Lovell, K Criscione, B Collins, T Prisinzano, S Poulsen, M McLeish, G Grunewald, and J Martin. Fragment-based screening by x-ray crystallography, ms and isothermal titration calorimetry to identify pnmt (phenylethanolamine n-methyltransferase) inhibitors. *Biochem. J.*, 431:51–61, 2010.
 - [147] Federico Riccardi Sirtori, Dannica Caronni, Maristella Colombo, Claudio Dalvit, Mauro Paolucci, Luca Regazzoni, Carlo Visco, and Gianpaolo Fogliatto. Establish an automated flow injection esi-ms method for the screening of fragment based libraries: Application to hsp90. *European Journal of Pharmaceutical Sciences*, 76: 83–94, 2015.
 - [148] P.J. Hajduk, G. Shepperd, D.G. Nettesheim, E.T. Olejniczak, S.B. Shuker, R.P. Meadows, and S. W. Fesik. Discovery of potent nonpeptide inhibitors of stromelysin using SAR by NMR. *J. Am. Chem. Soc.*, 119:5818–5827, 1997.
 - [149] Kurt Wuthrich. *NMR of proteins and nucleic acids*. Wiley, 1986.
 - [150] Maurizio Pellecchia, Daniel S Sem, and Kurt Wüthrich. Nmr in drug discovery. *Nat. Rev. Drug. Discov.*, 1(3):211–219, 2002.
 - [151] Maurizio Pellecchia, Ivano Bertini, David Cowburn, Claudio Dalvit, Ernest Giralt, Wolfgang Jahnke, Thomas L James, Steve W Homans, Horst Kessler, Claudio Luchinat, et al. Perspectives on nmr in drug discovery: a technique comes of age. *Nat. Rev. Drug. Discov.*, 7(9):738–745, 2008.
 - [152] Wolfgang Jahnke. Perspectives of biomolecular nmr in drug discovery: the blessing and curse of versatility. *Journal of biomolecular NMR*, 39(2):87–90, 2007.
 - [153] Brian J Stockman and Claudio Dalvit. Nmr screening techniques in drug discovery and drug design. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 41(3): 187–231, 2002.

- [154] Jochen Klages, Murray Coles, and Horst Kessler. Nmr-based screening: a powerful tool in fragment-based drug discovery. *Analyst*, 132(7):692–705, 2007.
- [155] Moriz Mayer and Bernd Meyer. Group epitope mapping by saturation transfer difference nmr to identify segments of a ligand in direct contact with a protein receptor. *J. Am. Chem. Soc.*, 123(25):6108–6117, 2001.
- [156] Moriz Mayer and Bernd Meyer. Characterization of ligand binding by saturation transfer difference nmr spectroscopy. 38(12):1784–1788, 1999.
- [157] Claudio Dalvit, Paolo Pevarello, Marco Tatò, Marina Veronesi, Anna Vulpetti, and Michael Sundström. Identification of compounds with binding affinity to proteins via magnetization transfer from bulk water*. *Journal of biomolecular NMR*, 18(1):65–68, 2000.
- [158] Claudio Dalvit, GianPaolo Fogliatto, Albert Stewart, Marina Veronesi, and Brian Stockman. Waterlogsy as a method for primary nmr screening: practical aspects and range of applicability. *Journal of biomolecular NMR*, 21(4):349–359, 2001.
- [159] Paul A Brough, Xavier Barril, Jenifer Borgognoni, Patrick Chene, Nicholas GM Davies, Ben Davis, Martin J Drysdale, Brian Dymock, Suzanne A Eccles, Carlos Garcia-Echeverria, et al. Combining hit identification strategies: fragment-based and in silico approaches to orally active 2-aminothieno [2, 3-d] pyrimidine inhibitors of the hsp90 molecular chaperone. *J. Med. Chem.*, 52(15):4794–4809, 2009.
- [160] Tilman Oltersdorf, Steven W Elmore, Alexander R Shoemaker, Robert C Armstrong, David J Augeri, Barbara A Belli, Milan Bruncko, Thomas L Deckwerth, Jurgen Dinges, Philip J Hajduk, et al. An inhibitor of bcl-2 family proteins induces regression of solid tumours. *Nature*, 435(7042):677–681, 2005.
- [161] Discovery of a novel warhead against β -secretase through fragment-based lead generation §.
- [162] Brian J Stockman, Michael Kothe, Darcy Kohls, Laura Weibley, Brendan J Connolly, Alissa L Sheils, Qing Cao, Alan C Cheng, Lily Yang, Ajith V Kamath, et al. Identification of allosteric pif-pocket ligands for pdk1 using nmr-based fragment screening and 1h-15n trosy experiments. *Chemical biology & drug design*, 73(2):179–188, 2009.
- [163] Jan Drenth. *X-Ray Crystallography*. Wiley Online Library, 2007.
- [164] Alan C Gibbs, Marta C Abad, Xuqing Zhang, Brett A Tounge, Francis A Lewandowski, Geoffrey T Struble, Weimei Sun, Zhihua Sui, and Lawrence C

- Kuo. Electron density guided fragment-based lead discovery of ketohexokinase inhibitors. *J. Med. Chem.*, 53(22):7979–7991, 2010.
- [165] Thomas G Davies and Marko Hyvönen. *Fragment-based drug discovery and X-ray crystallography*, volume 317. Springer Science & Business Media, 2012.
- [166] Tom L Blundell, Harren Jhoti, and Chris Abell. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug. Discov.*, 1(1):45–54, 2002.
- [167] Roderick E Hubbard. Structure-based drug discovery and protein targets in the cns. *Neuropharmacology*, 60(1):7–23, 2011.
- [168] Tom L Blundell and S Patel. High-throughput x-ray crystallography for drug discovery. *Current opinion in pharmacology*, 4(5):490–496, 2004.
- [169] Christopher W Murray and Tom L Blundell. Structural biology in fragment-based drug design. *Current opinion in structural biology*, 20(4):497–507, 2010.
- [170] Paul G Wyatt, Andrew J Woodhead, Valerio Berdini, John A Boulstridge, Maria G Carr, David M Cross, Deborah J Davis, Lindsay A Devine, Theresa R Early, Ruth E Feltell, et al. Identification of n-(4-piperidinyl)-4-(2, 6-dichlorobenzoylamino)-1 h-pyrazole-3-carboxamide (at7519), a novel cyclin dependent kinase inhibitor using fragment-based x-ray crystallography and structure based drug design†. *J. Med. Chem.*, 51(16):4986–4999, 2008.
- [171] Steven Howard, Valerio Berdini, John A Boulstridge, Maria G Carr, David M Cross, Jayne Curry, Lindsay A Devine, Theresa R Early, Lynsey Fazal, Adrian L Gill, et al. Fragment-based discovery of the pyrazol-4-yl urea (at9283), a multi-targeted kinase inhibitor with potent aurora kinase activity†. *J. Med. Chem.*, 52(2):379–388, 2008.
- [172] Stephen Antonysamy, Gavin Hirst, Frances Park, Paul Sprengeler, Frank Stappenbeck, Ruo Steensma, Mark Wilson, and Melissa Wong. Fragment-based discovery of jak-2 inhibitors. *Bioorg. Med. Chem. Lett.*, 19(1):279–282, 2009.
- [173] Michele N Schulz, Joerg Fanghaenel, Martina Schaefer, Volker Badock, Hans Briem, Ulf Boemer, Duy Nguyen, Manfred Husemann, and Roman C Hillig. A crystallographic fragment screen identifies cinnamic acid derivatives as starting points for potent pim-1 inhibitors. *Acta Crystallographica Section D: Biological Crystallography*, 67(3):156–166, 2011.
- [174] Nigel Howard, Chris Abell, Wendy Blakemore, Gianni Chessari, Miles Congreve, Steven Howard, Harren Jhoti, Christopher W Murray, Lisa CA Seavers, and

- Rob LM van Montfort. Application of fragment screening and fragment linking to the discovery of novel thrombin inhibitors. *J. Med. Chem.*, 49(4):1346–1355, 2006.
- [175] Paul CD Hawkins, A Geoffrey Skillman, and Anthony Nicholls. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.*, 50(1):74–82, 2007.
- [176] NM OpenEye Scientific Software, Santa Fe. *ROCS* 3.2.1.4, . URL <http://www.eyesopen.com>. (accessed Nov 15, 2014).
- [177] Alvin W. Hung, H. Leonardo Silvestre, Shijun Wen, Alessio Ciulli, Tom L. Blundell, and Chris Abell. Application of fragment growing and fragment linking to the discovery of inhibitors of mycobacterium tuberculosis pantothenate synthetase. *Angew. Chem. Int. Ed.*, 48(45):8452–8456, 2009.
- [178] Suhman Chung, Jared B Parker, Mario Bianchet, L Mario Amzel, and James T Stivers. Impact of linker strain and flexibility in the design of a fragment-based inhibitor. *Nat. Chem. Biol.*, 5:407–413, 2009.
- [179] Gerdien E de Kloe, David Bailey, Rob Leurs, and Iwan JP de Esch. Transforming fragments into candidates: small becomes big in medicinal chemistry. *Drug Discov. Today*, 14(13):630–646, 2009.
- [180] Miles Congreve, Gianni Chessari, Dominic Tisi, and Andrew J Woodhead. Recent developments in fragment-based drug discovery. *J. Med. Chem.*, 51(13):3661–3680, 2008.
- [181] Gianni Chessari and Andrew J Woodhead. From fragment to clinical candidate—a historical perspective. *Drug Discov. Today*, 14(13):668–675, 2009.
- [182] Yuan Cheng, Ted C Judd, Michael D Bartberger, James Brown, Kui Chen, Robert T Freneau Jr, Dean Hickman, Stephen A Hitchcock, Brad Jordan, Vivian Li, et al. From fragment screening to in vivo efficacy: optimization of a series of 2-aminoquinolines as potent inhibitors of beta-site amyloid precursor protein cleaving enzyme 1 (bace1). *J. Med. Chem.*, 54(16):5836–5857, 2011.
- [183] Steven J Taylor, Asitha Abeywardane, Shuang Liang, Ingo Muegge, Anil K Padyana, Zhaoming Xiong, Melissa Hill-Drzewi, Bennett Farmer, Xiang Li, Brandon Collins, et al. Fragment-based discovery of indole inhibitors of matrix metalloproteinase-13. *J. Med. Chem.*, 54(23):8174–8187, 2011.
- [184] Samantha J Hughes, David S Millan, Iain C Kilty, Russell A Lewthwaite, John P Mathias, Mark A O’Reilly, Andrew Pannifer, Anne Phelan, Frank Stühmeier,

- Darren A Baldock, et al. Fragment based discovery of a novel and selective pi3 kinase inhibitor. *Bioorg. Med. Chem. Lett.*, 21(21):6586–6590, 2011.
- [185] Valerie J. Gillet, Glenn Myatt, Zsolt Zsoldos, and a. Peter Johnson. Sprout, hippo and caesa: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspectives in Drug Discovery and Design*, 3(1):34–50, 1995.
- [186] David a. Pearlman and Mark a. Murcko. Concerts: Dynamic connection of fragments as an approach to de novo ligand design. *J. Med. Chem.*, 39(8):1651–1663, 1996.
- [187] Patrick Maass, Tanja Schulz-Gasch, Martin Stahl, and Matthias Rarey. Recore: A fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *J. Chem. Inf. Model.*, 47(2):390–399, 2007.
- [188] Georges Lauri and Paul a. Bartlett. CAVEAT: A program to facilitate the design of organic molecules. *J. Comput.-Aided Mol. Design*, 8(1):51–66, 1994.
- [189] Albert C. Pierce, Govinda Rao, and Guy W. Bemis. Breed: Generating novel inhibitors through hybridization of known ligands. application to cdk2, p38, and hiv protease. *J. Med. Chem.*, 47(11):2768–2775, 2004.
- [190] NM OpenEye Scientific Software, Santa Fe. *brood* 1.6.3.1, . URL <http://www.eyesopen.com>. (accessed Nov 15, 2014).
- [191] Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *Curr. Med. Chem.*, 3(10):1503–1507, 2008. doi: 10.1002/cmdc.200800178.
- [192] Xiao Qing Lewell, Duncan B Judd, Stephen P Watson, and Michael M Hann. Recap retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.*, 38(3):511–522, 1998.
- [193] Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.*, 1(1):1–11, 2009.
- [194] Sean A Hudson, Sachin Surade, Anthony G Coyne, Kirsty J McLean, David Leys, Andrew W Munro, and Chris Abell. Overcoming the limitations of fragment merging: rescuing a strained merged fragment series targeting mycobacterium tuberculosis cyp121. *ChemMedChem*, 8(9):1451–1456, 2013.
- [195] Jasna Fejzo, Christopher Lepre, and Xiaoling Xie. Application of nmr screening in drug discovery. *Curr. Top. Med. Chem.*, 3(1):81–97, 2003.

- [196] William P Jencks. On the attribution and additivity of binding energies. *Proc. Natl. Acad. Sci. U. S. A.*, 78(7):4046–4050, 1981.
- [197] Anna Kohlmann, Stephan G Zech, Feng Li, Tianjun Zhou, Rachel M Squillace, Lois Commodore, Matthew T Greenfield, Xiaohui Lu, David P Miller, Wei-Sheng Huang, et al. Fragment growing and linking lead to novel nanomolar lactate dehydrogenase inhibitors. *J. Med. Chem.*, 56(3):1023–1040, 2013.
- [198] Jeffrey R Huth, Chang Park, Andrew M Petros, Aaron R Kunzer, Michael D Wendt, Xilu Wang, Christopher L Lynch, Jamey C Mack, Kerry M Swift, Russell A Judge, et al. Discovery and design of novel hsp90 inhibitors using multiple fragment-based design strategies. *Chemical biology & drug design*, 70(1):1–12, 2007.
- [199] Michèle N Schulz and Roderick E Hubbard. Recent progress in fragment-based lead discovery. *Current opinion in pharmacology*, 9(5):615–621, 2009.
- [200] Wolfgang Jahnke, Jean-Michel Rondeau, Simona Cotesta, Andreas Marzinzik, Xavier Pellé, Martin Geiser, André Strauss, Marjo Götte, Francis Bitsch, René Hemmig, et al. Allosteric non-bisphosphonate fpps inhibitors identified by fragment-based discovery. *Nat. Chem. Biol.*, 6(9):660–666, 2010.
- [201] Daniel A Erlanson, Stephen W Fesik, Roderick E Hubbard, Wolfgang Jahnke, and Harren Jhoti. Twenty years on: the impact of fragments on drug discovery. *Nature Reviews Drug Discovery*, 2016.
- [202] David Filmore. It’s a GPCR world. *Modern Drug Discovery*, 7:24–28, 2004.
- [203] JP Overington, B Al-Lazikani, and Andrew L Hopkins. How many drug targets are there? *Nat. Rev. Drug. Discov.*, 5(12):993–996, 2006.
- [204] Eyal Vardy and Bryan L Roth. Conformational ensembles in gpcr activation. *Cell*, 152(3):385–386, 2013.
- [205] Ru Zhang and Xin Xie. Tools for gpcr drug discovery. *Acta Pharmacol. Sin.*, 33(3):372–384, 2012.
- [206] Graeme Milligan. Principles: extending the utility of [35 s] gtpγs binding assays. *Trends Pharmacol. Sci.*, 24(2):87–90, 2003.
- [207] Christine Williams. camp detection methods in hts: selecting the best from the rest. *Nat. Rev. Drug. Discov.*, 3(2):125–135, 2004.
- [208] Michael J Berridge and CW Taylor. Inositol trisphosphate and calcium signaling. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 53, pages 927–933. Cold Spring Harbor Laboratory Press, 1988.

- [209] Ali Salahpour and Larry S Barak. Visualizing receptor endocytosis and trafficking. *Signal Transduction Protocols*, pages 311–323, 2011.
- [210] Miranda MC van Der Lee, Maaïke Bras, Chris J Van Koppen, and Guido JR Zaman. β -arrestin recruitment assay for the identification of agonists of the sphingosine 1-phosphate receptor edg1. *J. biomol. screen.*, 2008.
- [211] Jonathan D. Violin and Robert J. Lefkowitz. β -arrestin-biased ligands at seven-transmembrane receptors. *Trends Pharmacol. Sci.*, 28(8):416–422, 2007. doi: <http://dx.doi.org/10.1016/j.tips.2007.06.006>. Special Issue on Allosterism and Collateral Efficacy.
- [212] Maria Martí-Solano, Denis Schmidt, Peter Kolb, and Jana Selent. Drugging specific conformational states of gpcrs: challenges and opportunities for computational chemistry. *Drug Discov. Today*, 21(4):625–631, 2016.
- [213] Bruno H Meyer, Jean-Manuel Segura, Karen L Martinez, Ruud Hovius, Nathalie George, Kai Johnsson, and Horst Vogel. FRET imaging reveals that functional neurokinin-1 receptors are monomeric and reside in membrane microdomains of live cells. *Proc. Natl. Acad. Sci. U. S. A.*, 103(7):2138–2143, 2006.
- [214] David W Piston and Gert-Jan Kremers. Fluorescent protein FRET: the good, the bad and the ugly. *Trends Biochem. Sci.*, 32(9):407–414, 2007.
- [215] Martin Cottet, Laura Albizu, Laetitia Comps-Agrar, Eric Trinquet, Jean-Philippe Pin, Bernard Mouillac, and Thierry Durroux. Time resolved FRET strategy with fluorescent ligands to analyze receptor interactions in native tissues: application to GPCR oligomerization. In *Receptor Signal Transduction Protocols*, pages 373–387. Springer, 2011.
- [216] Damien Maurel, Laetitia Comps-Agrar, Carsten Brock, Marie-Laure Rives, Emmanuel Bourrier, Mohammed Akli Ayoub, Hervé Bazin, Norbert Tinel, Thierry Durroux, Laurent Prézeau, et al. Cell-surface protein-protein interaction analysis with time-resolved FRET and SNAP-tag technologies: application to GPCR oligomerization. *Nat. methods*, 5(6):561–567, 2008.
- [217] Raphael Rozenfeld and Lakshmi A Devi. Exploring a role for heteromerization in GPCR signalling specificity. *Biochem. J.*, 433(1):11–18, 2011.
- [218] Raphael Rozenfeld and Lakshmi A Devi. Receptor heteromerization and drug discovery. *Trends. Pharmacol. Sci.*, 31(3):124–130, 2010.
- [219] Peter Kolb and Gerhard Klebe. The golden age of GPCR structural biology: Any impact on drug design? *Angew. Chem. Int. Ed.*, 50(49):11573–11575, 2011. doi: [10.1002/anie.201105869](https://doi.org/10.1002/anie.201105869).

- [220] Vadim Cherezov, Daniel M. Rosenbaum, Michael A. Hanson, Søren G. F. Rasmussen, Foon Sun Thian, Tong Sun Kobilka, Hee-Jung Choi, Peter Kuhn, William I. Weis, Brian K. Kobilka, and Raymond C. Stevens. High-resolution crystal structure of an engineered human β_2 -adrenergic G protein-coupled receptor. *Science*, 318(5854):1258–1265, 2007. doi: {10.1126/science.1150577}.
- [221] Daniel M. Rosenbaum, Vadim Cherezov, Michael A. Hanson, Søren G. F. Rasmussen, Foon Sun Thian, Tong Sun Kobilka, Hee-Jung Choi, Xiao-Jie Yao, William I. Weis, Raymond C. Stevens, and Brian K. Kobilka. GPCR engineering yields high-resolution structural insights into β_2 -adrenergic receptor function. *Science*, 318(5854):1266–1273, 2007. doi: {10.1126/science.1150609}.
- [222] Peter Kolb, Daniel M. Rosenbaum, John J. Irwin, Juan Jose Fung, Brian K. Kobilka, and Brian K. Shoichet. Structure-based discovery of β_2 -adrenergic receptor ligands. *Proc. Natl. Acad. Sci. U. S. A.*, 106(16):6843–6848, APR 21 2009. doi: {10.1073/pnas.0812657106}.
- [223] Daniel Wacker, Gustavo Fenalti, Monica A. Brown, Vsevolod Katritch, Ruben Abagyan, Vadim Cherezov, and Raymond C. Stevens. Conserved binding mode of human β_2 adrenergic receptor inverse agonists and antagonist revealed by X-ray crystallography. *J. Am. Chem. Soc.*, 132(33):11443–11445, 2010. doi: 10.1021/ja105108q.
- [224] Soren G. F. Rasmussen, Brian T. DeVree, Yaozhong Zou, Andrew C. Kruse, Ka Young Chung, Tong Sun Kobilka, Foon Sun Thian, Pil Seok Chae, Els Parndon, Diane Calinski, Jesper M. Mathiesen, Syed T. A. Shah, Joseph A. Lyons, Martin Caffrey, Samuel H. Gellman, Jan Steyaert, Georgios Skinotis, William I. Weis, Roger K. Sunahara, and Brian K. Kobilka. Crystal structure of the beta(2) adrenergic receptor-Gs protein complex. *Nature*, 477(7366):549–U311, 2011. doi: {10.1038/nature10361}.
- [225] Daniel M Rosenbaum, Cheng Zhang, Joseph A Lyons, Ralph Holl, David Aragao, Daniel H Arlow, Søren GF Rasmussen, Hee-Jung Choi, Brian T DeVree, Roger K Sunahara, et al. Structure and function of an irreversible agonist-[bgr] 2 adrenoceptor complex. *Nature*, 469(7329):236–240, 2011.
- [226] Aaron M Ring, Aashish Manglik, Andrew C Kruse, Michael D Enos, William I Weis, K Christopher Garcia, and Brian K Kobilka. Adrenaline-activated structure of β_2 -adrenoceptor stabilized by an engineered nanobody. *Nature*, 502(7472):575–9, October 2013.
- [227] Dietmar Weichert, Andrew C Kruse, Aashish Manglik, Christine Hiller, Cheng Zhang, Harald Hübner, Brian K Kobilka, and Peter Gmeiner. Covalent agonists

- for studying g protein-coupled receptor activation. *Proc. Natl. Acad. Sci. U. S. A.*, 111(29):10744–10748, 2014.
- [228] Søren GF Rasmussen, Hee-Jung Choi, Juan Jose Fung, Els Pardon, Paola Casarosa, Pil Seok Chae, Brian T DeVree, Daniel M Rosenbaum, Foon Sun Thian, Tong Sun Kobilka, A Schnapp, I Konetzki, Sunahara Roger K, SH Gellman, A Pautsch, J Steyaert, WI Weis, and Brian K Kobilka. Structure of a nanobody-stabilized active state of the beta(2) adrenoceptor. *Nature*, 469(7329):175–180, 2011.
- [229] Tony Warne, Maria J. Serrano-Vega, Jillian G. Baker, Rouslan Moukhametzianov, Patricia C. Edwards, Richard Henderson, Andrew G. W. Leslie, Christopher G. Tate, and Gebhard F. X. Schertler. Structure of a β_1 -adrenergic G-protein-coupled receptor. *Nature*, 454(7203):486–U2, 2008. doi: {10.1038/nature07101}.
- [230] Guillaume Lebon, Tony Warne, Patricia C Edwards, Kirstie Bennett, Christopher J Langmead, Andrew GW Leslie, and Christopher G Tate. Agonist-bound adenosine a2a receptor structures reveal common features of gpcr activation. *Nature*, 474(7352):521–525, 2011.
- [231] Ellen YT Chien, Wei Liu, Qiang Zhao, Vsevolod Katritch, Gye Won Han, Michael A Hanson, Lei Shi, Amy Hauck Newman, Jonathan A Javitch, Vadim Cherezov, et al. Structure of the human dopamine d3 receptor in complex with a d2/d3 selective antagonist. *Science*, 330(6007):1091–1095, 2010.
- [232] Beili Wu, Ellen Y. T. Chien, Clifford D. Mol, Gustavo Fenalti, Wei Liu, Vsevolod Katritch, Ruben Abagyan, Alexei Brooun, Peter Wells, F. Christopher Bi, Damon J. Hamel, Peter Kuhn, Tracy M. Handel, Vadim Cherezov, and Raymond C. Stevens. Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science*, 330(6007):1066–1071, 2010. doi: 10.1126/science.1194396.
- [233] Tatsuro Shimamura, Mitsunori Shiroishi, Simone Weyand, Hirokazu Tsujimoto, Graeme Winter, Vsevolod Katritch, Ruben Abagyan, Vadim Cherezov, Wei Liu, Gye Won Han, et al. Structure of the human histamine h1 receptor complex with doxepin. *Nature*, 475(7354):65–70, 2011.
- [234] Kazuko Haga, Andrew C. Kruse, Hidetsugu Asada, Takami Yurugi-Kobayashi, Mitsunori Shiroishi, Cheng Zhang, William I. Weis, Tetsuji Okada, Brian K. Kobilka, Tatsuya Haga, and Takuya Kobayashi. Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature*, 482(7386):547–U147, 2012. doi: {10.1038/nature10753}.

- [235] Jie Yin, Juan Carlos Mobarec, Peter Kolb, and Daniel M. Rosenbaum. Crystal structure of the human OX2 orexin receptor bound to the insomnia drug suvorexant. *Nature*, 519(7542):247–250, 2015. doi: {10.1038/nature14035}.
- [236] Daniel A Erlanson. Introduction to fragment-based drug discovery. In *Fragment-based drug discovery and X-ray crystallography*, pages 1–32. Springer, 2011.
- [237] Tobias Fink and Jean-Louis Reymond. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discove. *J. Chem. Inf. Model.*, 47:342–353, 2007.
- [238] Injin Bang and Hee-Jung Choi. Structural features of β 2 adrenergic receptor: Crystal structures and beyond. *Molecules and cells*, 38(2):105, 2015.
- [239] Dennis W McGraw and Stephen B Liggett. Molecular mechanisms of β 2-adrenergic receptor function and regulation. *Proc. Natl. Acad. Sci. U. S. A.*, 2(4):292–296, 2005.
- [240] Stephen P Andrews, Giles A Brown, and John A Christopher. Structure-based and fragment-based gpcr drug discovery. *ChemMedChem*, 9(2):256–275, 2014.
- [241] Iva Navratilova, Jérémy Besnard, and Andrew L Hopkins. Screening for gpcr ligands using surface plasmon resonance. *ACS Med. Chem. Lett.*, 2(7):549–554, 2011.
- [242] John A Christopher, Jason Brown, Andrew S Dore, James C Errey, Markus Koglin, Fiona H Marshall, David G Myszk, Rebecca L Rich, Christopher G Tate, Benjamin Tehan, et al. Biophysical fragment screening of the β 1-adrenergic receptor: identification of high affinity arylpiperazine leads using structure-based drug design. *J. Med. Chem.*, 56(9):3446–3455, 2013.
- [243] Art D Bochevarov, Edward Harder, Thomas F Hughes, Jeremy R Greenwood, Dale A Braden, Dean M Philipp, David Rinaldo, Mathew D Halls, Jing Zhang, and Richard A Friesner. Jaguar: A high-performance quantum chemistry software program with strengths in life and materials sciences. *Int J Quantum Chem*, 113(18):2110–2142, 2013.
- [244] Juan A. Ballesteros and Harel Weinstein. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. In Stuart C. Sealfon, editor, *Receptor Molecular Biology*, volume 25 of *Methods in Neurosciences*, chapter 19, pages 366–428. Academic Press, San Diego, CA, 1995.

- [245] Brian K Kobilka. Amino and carboxyl terminal modifications to facilitate the production and purification of a g protein-coupled receptor. *Analytical biochemistry*, 231(1):269–271, 1995.
- [246] NM OpenEye Scientific Software, Santa Fe. *QUACPAC* 1.6.3.1, . URL <http://www.eyesopen.com>. (accessed Nov 15, 2014).
- [247] Gideon Bollag, Peter Hirth, James Tsai, Jiazhong Zhang, Prabha N Ibrahim, Hanna Cho, Wayne Spevak, Chao Zhang, Ying Zhang, Gaston Habets, Elizabeth a Burton, Bernice Wong, Garson Tsang, Brian L West, Ben Powell, Rafe Shellooe, Adhirai Marimuthu, Hoa Nguyen, Kam Y J Zhang, Dean R Artis, Joseph Schlessinger, Fei Su, Brian Higgins, Raman Iyer, Kurt D’Andrea, Astrid Koehler, Michael Stumm, Paul S Lin, Richard J Lee, Joseph Grippo, Igor Puzanov, Kevin B Kim, Antoni Ribas, Grant a McArthur, Jeffrey a Sosman, Paul B Chapman, Keith T Flaherty, Xiaowei Xu, Katherine L Nathanson, and Keith Nolop. Clinical efficacy of a raf inhibitor needs broad target blockade in braf-mutant melanoma. *Nature*, 467(7315):596–599, 2010.
- [248] Stephen D. Roughley and Allan M. Jordan. The medicinal chemist’s toolbox: An analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.*, 54(10):3451–3479, 2011.
- [249] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.
- [250] Frank A. Momany and Rebecca Rone. Validation of the general purpose QUANTA 3.2/CHARMm force field. *J. Comput. Chem.*, 13:888–900, 1992.
- [251] K.T. No, J.A. Grant, and H.A. Scheraga. Determination of net atomic charges using a modified partial equalization of orbital electronegativity method. 1. Application to neutral molecules as models for polypeptides. *J. Phys. Chem.*, 94:4732–4739, 1990.
- [252] K.T. No, J.A. Grant, M.S. Jhon, and H.A. Scheraga. Determination of net atomic charges using a modified partial equalization of orbital electronegativity method. 2. Application to ionic and aromatic molecules as models for polypeptides. *J. Phys. Chem.*, 94:4740–4746, 1990.
- [253] John J. Irwin and Brian K. Shoichet. ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, 45:177–182, 2005.
- [254] NM OpenEye Scientific Software, Santa Fe. *OMEGA* 2.5.1.4, . URL <http://www.eyesopen.com>. (accessed Nov 15, 2014).

- [255] John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.*, 52(7):1757–1768, 2012. doi: 10.1021/ci3001277.
- [256] J. Steyaert, T. Laeremans, and E. Pardon. Novel chimeric polypeptides for screening and drug discovery purposes, August 7 2014. WO Patent App. PC-T/EP2014/051,845.
- [257] Greg Landrum. RDKit: Open-source chemoinformatics. URL <http://www.rdkit.org>. (accessed Nov 15, 2014).
- [258] Kerim Babaoglu and Brian K. Shoichet. Deconstructing fragment-based inhibitor discovery. *Nat. Chem. Biol.*, 2:720–723, 2006.
- [259] Peter Kolb and John J. Irwin. Docking screens: right for the right reasons? *Curr. Top. Med. Chem.*, 9:755–770, 2009.
- [260] Peter Kolb, Rafaela S. Ferreira, John J. Irwin, and Brian K. Shoichet. Docking & chemoinformatic screens for new ligands and targets. *Curr. Opin. Biotechnol.*, 20:429–436, 2009.
- [261] Florent Chevillard and Peter Kolb. SCUBIDOO: A large yet screenable and easily searchable database of computationally created chemical compounds optimized toward high likelihood of synthetic tractability. *J. Chem. Inf. Model.*, 55(9):1824–1835, 2015. doi: 10.1021/acs.jcim.5b00203.
- [262] Laurent Brault, Christelle Gasser, Franz Bracher, Kilian Huber, Stefan Knapp, and Jürg Schwaller. Pim serine/threonine kinases in the pathogenesis and therapy of hematologic malignancies and solid cancers. *haematologica*, 95(6):1004–1015, 2010.
- [263] Malte Bachmann, Christian Kosan, Pei Xiang Xing, Mathias Montenarh, Ingrid Hoffmann, and Tarik Möröy. The oncogenic serine/threonine kinase pim-1 directly phosphorylates and activates the g2/m specific phosphatase cdc25c. *Int. J. Biochem. Cell Biol.*, 38(3):430–443, 2006.
- [264] Robert Amson, Francois Sigaux, Serge Przedborski, Georges Flandrin, David Givol, and Adam Telerman. The human protooncogene product p33pim is expressed during fetal hematopoiesis and in diverse leukemias. *pnas*, 86(22):8857–8861, 1989.
- [265] Wei Wei Chen, Daniel C Chan, Carlton Donald, Michael B Lilly, and Andrew S Kraft. Pim family kinases enhance tumor growth of prostate cancer cells. *Mol. Cancer Res.*, 3(8):443–451, 2005.

- [266] Hong Dai, Rile Li, Thomas Wheeler, Andrea Diaz de Vivar, Anna Frolov, Salahdin Tahir, Irina AgoulNIK, Timothy Thompson, David Rowley, and Gustavo Ayala. Pim-2 upregulation: Biological implications associated with disease progression and perinueral invasion in prostate cancer. *Prostate*, 65(3):276–286, 2005.
- [267] Ying-Yi Li, Boryana K Popivanova, Yuichiro Nagai, Hiroshi Ishikura, Chifumi Fujii, and Naofumi Mukaida. Pim-3, a proto-oncogene with serine/threonine kinase activity, is aberrantly expressed in human pancreatic cancer and phosphorylates bad to block bad-mediated apoptosis in human pancreatic cancer cell lines. *Cancer Res.*, 66(13):6741–6747, 2006.
- [268] Boryana Konstantinova Popivanova, Ying-Yi Li, Huachuan Zheng, Kenji Omura, Chifumi Fujii, Koichi Tsuneyama, and Naofumi Mukaida. Proto-oncogene, pim-3 with serine/threonine kinase activity, is aberrantly expressed in human colon cancer cells and can prevent bad-mediated apoptosis. *Cancer Sci.*, 98(3):321–328, 2007.
- [269] Martijn C Nawijn, Andrej Alendar, and Anton Berns. For better or for worse: the role of pim oncogenes in tumorigenesis. *Nature Reviews Cancer*, 11(1):23–34, 2011.
- [270] Albert C Pierce, Marc Jacobs, and Cameron Stuver-Moody. Docking study yields four novel inhibitors of the protooncogene pim-1 kinase. *J. Med. Chem.*, 51(6):1972–1975, 2008.
- [271] Alexey Ishchenko, Lin Zhang, Jean-Yves Le Brazidec, Junhua Fan, Jer Hong Chong, Aparna Hingway, Annie Raditsis, Latika Singh, Brian Elenbaas, Victor Sukbong Hong, et al. Structure-based design of low-nanomolar pim kinase inhibitors. *Bioorg. Med. Chem. Lett.*, 25(3):474–480, 2015.
- [272] Matthew Burger, Mika Lindvall, HAN Wooseok, Jiong Lan, Gisele Nishiguchi, Cynthia Shafer, Cornelia Bellamacina, Kay Huh, Gordana Atallah, Christopher McBride, et al. Pim kinase inhibitors and methods of their use, September 2014. US Patent 8,822,497.
- [273] Vanda Pogacic, Alex N Bullock, Oleg Fedorov, Panagis Filippakopoulos, Christelle Gasser, Andrea Biondi, Sandrine Meyer-Monard, Stefan Knapp, and Juerg Schwaller. Structural analysis identifies imidazo [1, 2-b] pyridazines as pim kinase inhibitors with in vitro antileukemic activity. *Cancer research*, 67(14):6916–6924, 2007.

- [274] Cynthia M Shafer, Mika Lindvall, Cornelia Bellamacina, Thomas G Gesner, Asha Yabannavar, Weiping Jia, Song Lin, and Annette Walter. 4-(1h-indazol-5-yl)-6-phenylpyrimidin-2 (1h)-one analogs as potent cdc7 inhibitors. *Bioorg. Med. Chem. Lett.*, 18(16):4482–4485, 2008.
- [275] Joaquín Pastor, Julen Oyarzabal, Gustavo Saluste, Rosa María Alvarez, Virginia Rivero, Francisco Ramos, Elena Cendón, Carmen Blanco-Aparicio, Nuria Ajenjo, Antonio Cebriá, et al. Hit to lead evaluation of 1, 2, 3-triazolo [4, 5-b] pyridines as pim kinase inhibitors. *Bioorg. Med. Chem. Lett.*, 22(4):1591–1597, 2012.
- [276] Marc D Jacobs, James Black, Olga Futer, Lora Swenson, Brian Hare, Mark Fleming, and Kumkum Saxena. Pim-1 ligand-bound structures reveal the mechanism of serine/threonine kinase inhibition by ly294002. *J. Biol. Chem.*, 280(14):13728–13734, 2005.
- [277] Alex N Bullock, Judit Debreczeni, Ann L Amos, Stefan Knapp, and Benjamin E Turk. Structure and substrate specificity of the pim-1 kinase. *J. Biol. Chem.*, 280(50):41675–41682, 2005.
- [278] Kevin Qian, Lian Wang, Charles L Cywin, Bennett T Farmer, Eugene Hickey, Carol Homon, Scott Jakes, Mohammed A Kashem, George Lee, Scott Leonard, et al. Hit to lead account of the discovery of a new class of inhibitors of pim kinases and crystallographic studies revealing an unusual kinase binding mode†. *jmc*, 52(7):1814–1827, 2009.
- [279] Andrew C Good, Jinyu Liu, Bradford Hirth, Gary Asmussen, Yibin Xiang, Hans-Peter Biemann, Kimberly A Bishop, Trisha Fremgen, Maria Fitzgerald, Tatiana Gladysheva, et al. Implications of promiscuous pim-1 kinase fragment inhibitor hydrophobic interactions for fragment-based drug design. *J. Med. Chem.*, 55(6):2641–2648, 2012.
- [280] Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40(D1):D1100–D1107, 2012. doi: 10.1093/nar/gkr777.
- [281] Michael M Mysinger, Michael Carchia, John J Irwin, and Brian K Shoichet. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. *J. Med. Chem.*, 55(14):6582–6594, 2012.
- [282] Junqi Li, Steven G Ballmer, Eric P Gillis, Seiko Fujii, Michael J Schmidt, Andrea M E Palazzolo, Jonathan W Lehmann, Greg F Morehouse, and Martin D Burke.

- Synthesis of many different types of organic small molecules using one automated process. *Science*, 347(6227):1221–1226, 2015.
- [283] ChemAxon. *Calculator Plugins* 16.1.11. URL <http://www.chemaxon.com>. (accessed Jan 11, 2016).
- [284] Gilles Klopman, Ju-Yun Li, Shaomeng Wang, and Mario Dimayuga. Computer automated log p calculations based on an extended group contribution approach. *J. Chem. Inf. Model.*, 34(4):752–781, 1994.
- [285] Edgar A Gatica and Claudio N Cavasotto. Ligand and decoy sets for docking to g protein-coupled receptors. *J. Chem. Inf. Model.*, 52(1):1–6, 2011.
- [286] Steven H. Bertz. The first general index of molecular complexity. *J. Am. Chem. Soc.*, 103:3599–3601, 1980.
- [287] Eric M Gordon, Ronald W Barrett, William J Dower, Stephen PA Fodor, and Mark A Gallop. Applications of combinatorial technologies to drug discovery. 2. combinatorial organic synthesis, library screening strategies, and future directions. *J. Med. Chem.*, 37(10):1385–1401, 1994.
- [288] Friedhelm Balkenhohl, Christoph von dem Bussche-Hünnefeld, Annegret Lansky, and Christian Zechel. Combinatorial synthesis of small organic molecules. *Angew. Chem. Int. Ed. Engl.*, 35(20):2288–2337, 1996.
- [289] Nicholas K Terrett, Mark Gardner, David W Gordon, Ryszard J Kobylecki, and John Steele. Combinatorial synthesis—the design of compound libraries and their application to drug discovery. *Tetrahedron*, 51(30):8135–8173, 1995.
- [290] John Carroll. Will combinatorial chemistry keep its promise? *Biotechnol. Healthc.*, 2(3):26, 2005.
- [291] Robert F. Service. The Synthesis Machine. *Science*, 347(6227):1190–1193, 2015.
- [292] Anton J Hopfinger, Emilio Xavier Esposito, Antonio Llinàs, Robert C Glen, and Jonathan M Goodman. Findings of the challenge to predict aqueous solubility. *J. Chem. Inf. Model.*, 49(1):1–5, 2008.
- [293] Antonio Llinàs, Robert C Glen, and Jonathan M Goodman. Solubility challenge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements? *J. Chem. Inf. Model.*, 48(7):1289–1303, 2008.

Erklärung

Ich versichere, dass ich meine Dissertation

”Improved approaches to ligand growing through fragment docking and fragment-based library design”

selbständig ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen bedient habe. Alle vollständig oder sinngemäß übernommenen sind Zitate als solche gekennzeichnet.

Die Dissertation wurde in der jetzigen oder einer ähnlichen Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen sonstigen Prüfungszwecken gedient.

Marburg, den 26/7/2016

(Florent Chevillard)

Florent Chevillard | CV

Wehrdaer Str. 54 – 35041 Marburg – Germany

☎ (DE) +49 1520 45 58093 • ☎ (FR) +33 6 99 92 15 12

☎ (DE) +49 6421 28 21351 • ✉ florent.chevillard@gmail.com

🌐 <http://florent.chevillard.free.fr>

Chemoinformatician & Bioinformatician

Hope for the best and plan for the worst

Personal

Date of birth: 05/08/1986

Place of birth: Toulouse, France

Nationality: French

Experience

Work.....

Chemoinformatics & FBDD, Lab of Dr. P. Kolb

Phillips-University Marburg

PhD Student

11/2011 - Present

My PhD studies focus on developing new *in silico* tools in order to assist fragment-based drug discovery. I work on two important drug targets, the β_2 adrenergic receptor and PIM-1 kinase.

Molecules Therapeutiques In Silico, Lab of Prof. B. Villoutreix

Paris Diderot University

Engineer

10/2010 - 10/2011

Prediction of different ADMET properties, development of chemical libraries and virtual screening (structure-based and ligand-based).

Internships.....

CERMN, Lab of Prof. R. Bureau

Caen Basse-Normandie University

Intern

02/2010 - 09/2010

QSAR / QSPR: prediction of the toxicity of small compounds.

Sanofi-Aventis, Structure Design & Informatics

Vitry-sur-Seine

Intern

06/2009 - 08/2009

Ligand-based approaches comparison for virtual screening.

Education

Chemoinformatics

Strasbourg

Louis Pasteur University

09/2008 - 09/2010

Master Degree

Organic & Inorganic Chemistry

Toulouse

Paul Sabatier University

09/2004 - 09/2008

Licence Degree

Languages

French: Fluent

Mother tongue

English: Fluent

TOEFL: 91

German: Casual

B1 level

Computer skills

Programming Languages.....

- Python
- php
- Javascript
- R-script
- HTML & CSS
- MySQL
- bash & tcsh scripting
- Pipeline Pilot Script

Frameworks.....

- bootstrap
- flask

Softwares.....

- Pipeline Pilot
- git
- R statistics
- Microsoft Office

Operating systems.....

- Linux
- Windows

Hardware.....

- Building & deploying computers and Home Theater Personal Computers (HTPC)

Publications

Published.....

- Chevillard F. and Kolb P. **SCUBIDOO: A large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized towards High Likelihood of Synthetic Tractability.** *J. Chem. Inf. Model.*, 55, 1824-1835. **2015.**
- Martiny V. Y., Carbonell P., Chevillard F., Moroy G., Nicot A. B., Vayer P., Villoutreix B. O., Miteva M. A. **Integrated structure- and ligand-based in silico approach to predict inhibition of cytochrome P450 2D6.** *Bioinformatics*, btv486. **2015.**
- Chevillard F., Lagorce D., Reynes C., Villoutreix B., Vayer P., Miteva M. **In Silico Prediction of Aqueous Solubility: A Multimodel Protocol Based on Chemical Similarity.** *Mol. Pharm.*, 9, 3127-3135. **2012.**

Submitted.....

- Gawron S., Aristotelous T., Chevillard F., Hopkins A. L., Kolb P., Navratilova I. H., Gilbert I. H. **Structure-free optimisation of fragments for the β_2 -adrenoreceptor.** **2016.**

To be submitted.....

- Chevillard F. *et al.* **Binding-site compatible growing applied to the design of β_2 adrenergic receptor ligands.** **2016.**

Talks

- **SCUBIDOO: a free online database exploring accessible chemical space.** 7th Chemoinformatic Days of the French Society of Chemoinformatics, Nice, France, **2015**. Best presentation award.
- **SCUBIDOO: a free online database exploring new chemical space.** Open Source Pharma, 2nd Annual Meeting, Rauischholzhausen, Germany, **2015**.

Posters

- Chevillard F., Kolb P. **SCUBIDOO: combinatorial synthesis applied to combinatorial growing.** Seventh Joint Sheffield Conference on Chemoinformatics, Sheffield, UK **2016**.
- Chevillard F., Kolb P. **SCUBIDOO: a free online database exploring accessible chemical space.** 11th German Conference on Chemoinformatics, Fulda, Germany, **2015**.
- Chevillard F., Kolb P. **De Novo Design of Selective Compounds: a Fragment Based Pipeline Applied to β_2 AR.** 9th German Conference on Chemoinformatics, Fulda, Germany, **2013**.
- Chevillard F., Kolb P. **Growing Specific Ligands for β_2 AR and HH1R from a Common Fragment-Sized Scaffold.** Summerschool of chemoinformatics, Obernai, France, **2012**.
- Chevillard F. **Beyond measuring the performance of ligand-based methods for back-screening selections.** Summerschool of chemoinformatics, Obernai, France, **2010**.

Workshops

- Chevillard F., Schmidt D. **Ligand docking in structure-based drug design.** GLISTEN Work group meeting and Workshop, Warsaw, Poland. **2013**.

Teaching

Introduction to Chemoinformatics

Philipps-University Marburg, Department of Chemistry
Semester 3

Marburg

2012–2015

Introduction to Pymol

Philipps-University Marburg, Department of Pharmacy
Semester 5

Marburg

2012–2015

Extracurricular Activity

Leader in Scouting Association

Scouts et Guides de France
Mentoring of the *Explorers* (ages 14 to 17)

Toulouse

09/2007–08/2008

Leader in Scouting Association

Scouts et Guides de France
Mentoring of the *Explorers* (ages 14 to 17)

Strasbourg

09/2008–08/2009

Acknowledgements

Research in the science field is a treacherous path, made of twists, turns and dead-ends. If it was not for some people I met during my career, I will probably not be sitting here, polishing this very manuscript. I wanted to take the time to acknowledge all the great scientists or human beings who helped me to reach this very point.

I guess I could say my scientific career started during the summer of 2009. I was my first internship and I had the opportunity to work at Sanofi-Aventis. It was a rich experience at many levels and I learned a lot during these three months. I would like to express my profound gratitude to Dr. Nicolas Baurin, my first mentor, who helped me to improve my skills as chemoinformatician, but also as a scientist. He taught me that I could do things that I thought I could not do. He is probably the one who ignited that sparkle which made the scientist I am today. I would also like to thank Dr. Herve Minoux and the entire drug design unit for the nice work atmosphere during this experience.

During the spring of 2010, I made a six months internship in order to complete my master degree in chemoinformatic. I went to the CERMN in Caen and worked under the supervision of Prof. Ronan Bureau. I would like to thank Ronan and Jana for their wonderful support, especially when the grant application for my potentially PhD did not work out. Their genuine concern touched me a lot. I would also like to thank my coworkers Sylvain, Alban and Fabienne for their unbeatable support and the positive work environment.

After a successful completion of my master degree, I had the chance to be hired at the lab of Prof. Bruno Villoutreix (MTi). It was my first real job and it resulted in a year full of new scientific experiences (my first two papers !). I would like to thank Prof. Maria Miteva and Bruno for their faith in me when writing a PhD grant, but also for their vigorous support when it did not work out. Even though it felt like an impression of "déjà vu", I was heavily surrounded (David, Julien, Olivier, Adrien and Leslie) and thanks to that (them), I was able to finally find what I was looking for: a position as graduate student.

I then went to Marburg at the lab of Dr Peter Kolb. New country, new world (protein: what ?), new tools (no more Pipeline Pilot...) and as expected, I acquired a lot of new abilities. I am naturally grateful to a lot of people, since I had the chance to be involved in many interdisciplinary collaborations.

I would like to thank our collaborators from Brussels who helped us for the application of PINGUI: Dr. Steven Ballet and Cecilia Betti for the synthesis part and Prof. Jan Steyaert and Dr. Els Pardon for conducting the assays.

I am grateful to Prof. Wibke Diederich and Helena Rimmer for trusting us and for conducting the synthesis of new products in two critical projects of my PhD.

I would like to express my deep gratitude to Prof. Gerhard Klebe, for co-supervising my PhD. I learned a lot from his pertinent questions during the seminars, and it helped me to improve over time, as scientist. I would like also to thank the entire AG Klebe and AG Kolb for the work environment full of expertise at many levels. Jonathan for his deep knowledge in chemistry and the help he provided me when conceiving SCUBIDOO. I had the chance to have a chemist who could tell me what informations he would like to see, and that was extremely helpful. Alex and Johannes for their divers and valuable scientific feedbacks. I like to thank Christof for his contribution to the PIM1 project (TSA and X-ray crystallography). I had the chance, as computational chemist, to work directly with chemists and biologists. This made my entire work more real and more exciting, since I had a direct follow-up of my computational predictions or applications.

I am grateful to Dr. Dimitrios Tzalis, current CEO of Taros, for trusting Peter and I, and help us to validate SCUBIDOO at a large (automatic) scale. I am fully aware of how lucky we are to have such collaboration. I would like to thank Dr. Anna Karawajczyk, for her expertise and support, as well as all the chemists who were involved in this project.

I thank Prof. Moritz Bünemann and Prof. Matthias Rarey for serving on my graduate examination committee and bringing their expertise (biology of GPCRs and chemoinformatic respectively) to my work.

I would like to thank the anonymous and numerous developers who contribute, every day, to open source softwares, such as Rdkit, bootstrap, R, Git and Linux in general.

From a non-science point of view, I would like to thank all my friends from soccer (*dozentenfußball*), Micha, Chris, Florian, Namir, Prof. Roland Hartmann. Wednesday was often the day I could not wait for, because I had (and still have) a lot of fun here.

I would like to express my genuine gratitude to my family and friends, who supported me during all along these years: my father, my mother (even though she does not really understand what I am doing), Violaine, Pierre-Yves, Sophie, Nicolas, Olivier, Caroline, Biou, Marie and my wonderful nephews and nieces. Christophe for his friendship during all those years. Jeremy for his scientific help at many key moments.

Last, but certainly not the least, I am deeply grateful to my supervisor Dr. Peter Kolb. It will be hard to list all the positive things he contributed to, but I will try to highlight the most important ones. If I had to summarize my PhD with one quote, I would chose the one from Winston Churchill: "*Success consists of going from failure to failure without loss of enthusiasm*". In my humble opinion, during my PhD, Peter contributed to at least 80% of the enthusiastic component. To put it simply, I couldn't have achieved what I have done without Peter's trust and support. There were many up and downs, and his enthusiasm during the downs was fantastic. I would say that one of the challenge I had, was to work on some side projects that were not as attracting as my main one. Thus, I invested less time (and focus) and ultimately these projects were more or less buried. It is part of the science process, you tried and you learned a lot of lessons, even though most of the bad results (that are still interesting) are not necessarily shown. From a personal point of view, this is still a rich experience since you also learn what really wakes your scientific curiosity (or not). I want to thank Peter for choosing not to push these projects, but rather encourage me to focus on what I would like to do. I guess it is also what a PhD is about: finding who you are or what you are good at. I had the chance to have a supervisor who trusted me and pretty much let me drive my projects in the last two years. This was probably the most exciting part of my PhD, since I felt I was exploring areas I was naturally curious about. I think the management from Peter payed off, since I am currently in a position where I can say I am working on what I like and want to keep exploring this very domain (i.e. post doc). This is also supported by the fact that we have many ongoing and yet-to-start exciting collaborations in the pipeline. I hope we will keep up the pace, but I am really looking forward the next couple of years. So thanks for that Peter.